SAS/STAT® 13.2 User's Guide

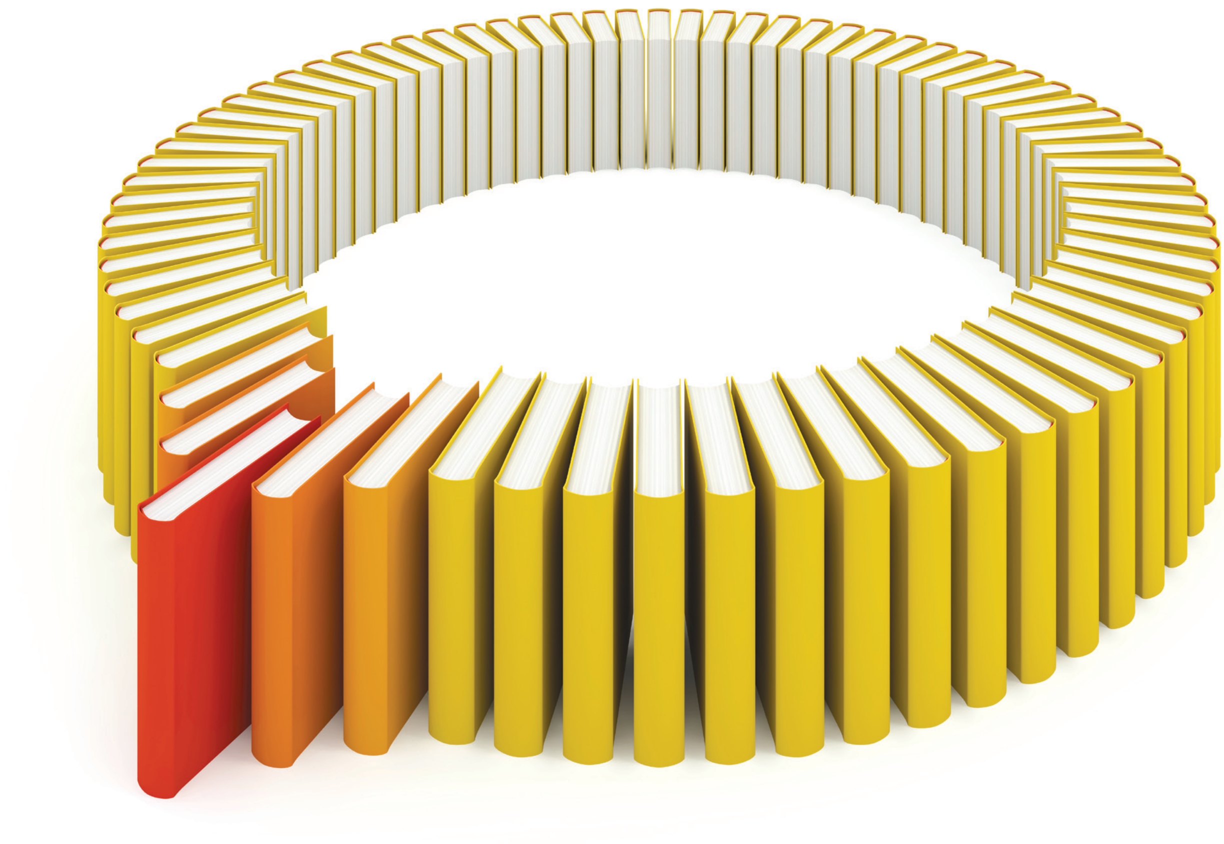
The SURVEYFREQ Procedure

This document is an individual chapter from *SAS/STAT® 13.2 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2014. *SAS/STAT® 13.2 User's Guide*. Cary, NC: SAS Institute Inc.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

August 2014



Chapter 97
The SURVEYFREQ Procedure

Contents

Overview: SURVEYFREQ Procedure

The SURVEYFREQ procedure produces one-way to n-way frequency and crosstabulation tables from sample survey data. These tables include estimates of population totals, population proportions, and their standard errors. Confidence limits, coefficients of variation, and design effects are also available. The procedure provides a variety of options to customize the table display.

For one-way frequency tables, PROC SURVEYFREQ provides Rao-Scott chi-square goodness-of-fit tests, which are adjusted for the sample design. You can test a null hypothesis of equal proportions for a one-way frequency table, or you can input custom nu5ll hypothesis proportions for the test. For two-way tables, PROC SURVEYFREQ provides design-adjusted tests of independence, or no association, between the row and column variables. These tests include the Rao-Scott chi-square test, the Rao-Scott likelihood ratio test, the Wald chi-square test, and the Wald log-linear chi-square test. For 2×2 tables, PROC SURVEYFREQ computes estimates and confidence limits for risks (row proportions), the risk difference, the odds ratio, and relative risks.

PROC SURVEYFREQ computes variance estimates based on the sample design used to obtain the survey data. The design can be a complex multistage survey design with stratification, clustering, and unequal weighting. PROC SURVEYFREQ provides a choice of variance estimation methods, which include Taylor series linearization, balanced repeated replication (BRR), and the jackknife.

PROC SURVEYFREQ uses ODS Graphics to create graphs as part of its output. For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." For specific information about the statistical graphics available with the SURVEYFREQ procedure, see the PLOTS= option in the TABLES statement and the section "ODS Graphics" on page 8046.

Getting Started: SURVEYFREQ Procedure

The following example shows how you can use PROC SURVEYFREQ to analyze sample survey data. The example uses data from a customer satisfaction survey for a student information system (SIS), which is a software product that provides modules for student registration, class scheduling, attendance, grade reporting, and other functions.

The software company conducted a survey of school personnel who use the SIS. A probability sample of SIS users was selected from the study population, which included SIS users at middle schools and high schools in the three-state area of Georgia, South Carolina, and North Carolina. The sample design for this survey was a two-stage stratified design. A first-stage sample of schools was selected from the list of schools in the three-state area that use the SIS. The list of schools (the first-stage sampling frame) was stratified by state and by customer status (whether the school was a new user of the system or a renewal user). Within the first-stage strata, schools were selected with probability proportional to size and with replacement, where the size measure was school enrollment. From each sample school, five staff members were randomly selected to complete the SIS satisfaction questionnaire. These staff members included three teachers and two administrators or guidance department members.

The SAS data set SIS_Survey contains the survey results, as well as the sample design information needed to analyze the data. This data set includes an observation for each school staff member responding to the survey. The variable Response contains the staff member's response about overall satisfaction with the system.

The variable State contains the school's state, and the variable NewUser contains the school's customer status ('New Customer' or 'Renewal Customer'). These two variables determine the first-stage strata from which schools were selected. The variable School contains the school identification code and identifies the first-stage sampling units (clusters). The variable SamplingWeight contains the overall sampling weight for each respondent. Overall sampling weights were computed from the selection probabilities at each stage of sampling and were adjusted for nonresponse.

Other variables in the data set SIS_Survey include SchoolType and Department. The variable SchoolType identifies the school as a high school or a middle school. The variable Department identifies the staff member as a teacher, or an administrator or guidance department member.

The following PROC SURVEYFREQ statements request a one-way frequency table for the variable Response:

```
title 'Student Information System Survey';
proc surveyfreq data=SIS_Survey;
   tables  Response;
   strata  State NewUser;
   cluster School;
   weight  SamplingWeight;
run;
```

The PROC SURVEYFREQ statement invokes the procedure and identifies the input data set to be analyzed. The TABLES statement requests a one-way frequency table for the variable Response. The table request syntax for PROC SURVEYFREQ is very similar to the table request syntax for PROC FREQ. This example shows a request for a single one-way table, but you can also request two-way tables and multiway tables. As in PROC FREQ, you can request more than one table in the same TABLES statement, and you can use multiple TABLES statements in the same invocation of the procedure.

The STRATA, CLUSTER, and WEIGHT statements provide sample design information for the procedure, so that the analysis is done according to the sample design used for the survey, and the estimates apply to the study population. The STRATA statement names the variables State and NewUser, which identify the first-stage strata. The design for this example also includes stratification at the second stage of selection (by type of school personnel), but you specify only the first-stage strata for PROC SURVEYFREQ. The CLUSTER statement names the variable School, which identifies the clusters (primary sampling units). The WEIGHT statement names the sampling weight variable.

Figure 97.1 and Figure 97.2 display the output produced by PROC SURVEYFREQ, which includes the "Data Summary" table and the one-way table, "Table of Response." The "Data Summary" table is produced by default unless you specify the NOSUMMARY option. This table shows there are 6 strata, 370 clusters or schools, and 1850 observations (respondents) in the SIS_Survey data set. The sum of the sampling weights is approximately 39,000, which estimates the total number of school personnel in the study area that use the SIS.

Figure 97.1 SIS_Survey Data Summary

Student Information System Survey

The SURVEYFREQ Procedure

Data Summary	
Number of Strata	6
Number of Clusters	370
Number of Observations	1850
Sum of Weights	38899.6482

Figure 97.2 displays the one-way table of Response, which provides estimates of the population total (weighted frequency) and the population percentage for each category (level) of the variable Response. The response level 'Very Unsatisfied' has a frequency of 304, which means that 304 sample respondents fall into this category. It is estimated that 17.17% of all school personnel in the study population fall into this category, and the standard error of this estimate is 1.29%. The estimates apply to the population of all SIS users in the study area, as opposed to describing only the sample of 1850 respondents. The estimate of the total number of school personnel that are 'Very Unsatisfied' is 6,678, with a standard deviation of 502. The standard errors computed by PROC SURVEYFREQ are based on the multistage stratified design of the survey. This differs from some of the traditional analysis procedures, which assume the design is simple random sampling from an infinite population.

Figure 97.2 One-Way Table of Response

Table of Response

Response	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
Very Unsatisfied	304	6678	501.61039	17.1676	1.2872
Unsatisfied	326	6907	495.94101	17.7564	1.2712
Neutral	581	12291	617.20147	31.5965	1.5795
Satisfied	455	9309	572.27868	23.9311	1.4761
Very Satisfied	184	3714	370.66577	9.5483	0.9523
Total	1850	38900	129.85268	100.000	

The following PROC SURVEYFREQ statements request confidence limits for the percentages, a chi-square goodness-of-fit test, and a weighted frequency plot for the one-way table of Response. The ODS GRAPHICS ON statement enables ODS Graphics.

```
title 'Student Information System Survey';
ods graphics on;
proc surveyfreq data=SIS_Survey nosummary;
   tables  Response / clwt nopct chisq
                      plots=WtFreqPlot;
   strata  State  NewUser;
   cluster School;
   weight  SamplingWeight;
run;
ods graphics off;
```

The NOSUMMARY option in the PROC SURVEYFREQ statement suppresses the "Data Summary" table. In the TABLES statement, the CLWT option requests confidence limits for the weighted frequencies (totals). The NOPCT option suppresses display of the weighted frequencies and their standard deviations. The CHISQ option requests a Rao-Scott chi-square goodness-of-fit test, and the PLOTS= option requests a weighted frequency plot. ODS Graphics must be enabled before producing plots.

Figure 97.3 shows the one-way table of Response, which includes confidence limits for the weighted frequencies. The 95% confidence limits for the total number of users that are 'Very Unsatisfied' are 5692 and 7665. You can change the confidence level by specifying the ALPHA= option; by default, ALPHA=0.05, which produces 95% confidence limits. Like the other estimates and standard errors produced by PROC SURVEYFREQ, these confidence limit computations take into account the complex survey design and apply to the entire study population.

Figure 97.3 Confidence Limits for Response Totals

Student Information System Survey

The SURVEYFREQ Procedure

Table of Response					
Response	**Frequency**	**Weighted Frequency**	**Std Dev of Wgt Freq**	**95% Confidence Limits for Wgt Freq**	
Very Unsatisfied	304	6678	501.61039	5692	7665
Unsatisfied	326	6907	495.94101	5932	7882
Neutral	581	12291	617.20147	11077	13505
Satisfied	455	9309	572.27868	8184	10435
Very Satisfied	184	3714	370.66577	2985	4443
Total	1850	38900	129.85268	38644	39155

Figure 97.4 displays the weighted frequency plot of Response. The plot displays weighted frequencies (totals) together with their confidence limits in the form of a vertical bar chart. You can use the PLOTS= option to request a dot plot instead of a bar chart or to plot percentages instead of weighted frequencies.

Figure 97.4 Bar Chart of Response Totals

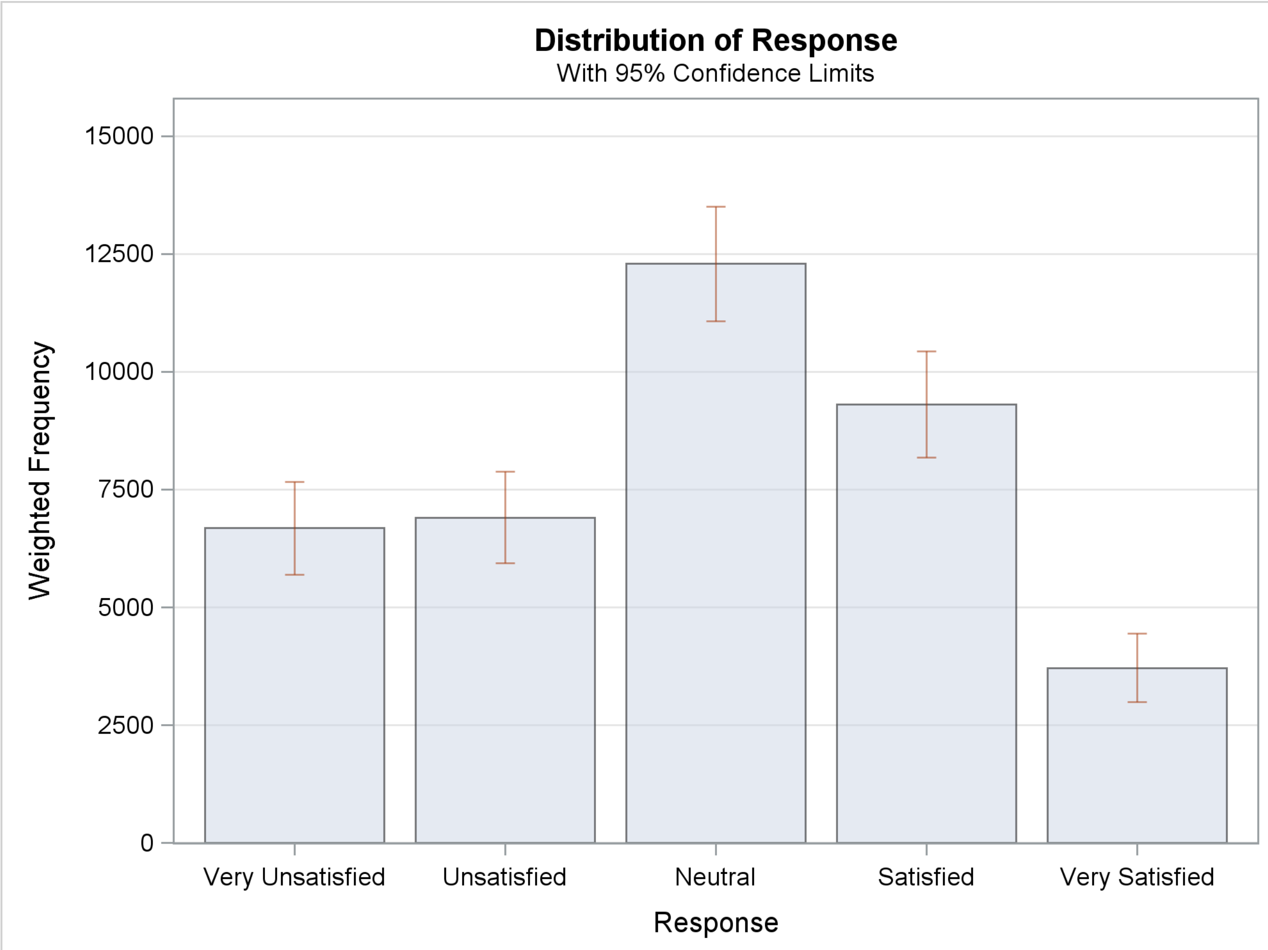


Figure 97.5 shows the chi-square goodness-of-fit results for the table of Response. The null hypothesis for this test is equal proportions for the levels of the one-way table. (To test a null hypothesis of specified proportions instead of equal proportions, you can use the TESTP= option to specify null hypothesis proportions.)

The chi-square test provided by the CHISQ option is the Rao-Scott design-adjusted chi-square test, which takes the sample design into account and provides inferences for the study population. To produce the Rao-Scott chi-square statistic, PROC SURVEYFREQ first computes the usual Pearson chi-square statistic based on the weighted frequencies, and then adjusts this value by using a design correction. An F approximation is also provided. For the table of Response, the F value is 30.0972 with a p-value of <0.0001, which indicates rejection of the null hypothesis of equal proportions for all response levels.

Figure 97.5 Chi-Square Goodness-of-Fit Test for Response

Rao-Scott Chi-Square Test	
Pearson Chi-Square	251.8105
Design Correction	2.0916
Rao-Scott Chi-Square	120.3889
DF	4
Pr > ChiSq	<.0001
F Value	30.0972
Num DF	4
Den DF	1456
Pr > F	<.0001
Sample Size = 1850	

Continuing to analyze the SIS_Survey data, the following PROC SURVEYFREQ statements request a two-way table of SchoolType by Response:

```
title 'Student Information System Survey';
ods graphics on;
proc surveyfreq data=SIS_Survey nosummary;
   tables  SchoolType * Response /
      plots=wtfreqplot(type=dot scale=percent groupby=row);
   strata  State NewUser;
   cluster School;
   weight  SamplingWeight;
run;
ods graphics off;
```

The STRATA, CLUSTER, and WEIGHT statements do not change from the one-way table analysis, because the sample design and the input data set are the same. These SURVEYFREQ statements request a different table but specify the same sample design information.

The ODS GRAPHICS ON statement enables ODS Graphics. The PLOTS= option in the TABLES statement requests a plot of SchoolType by Response, and the TYPE=DOT *plot-option* specifies a dot plot instead of the default bar chart. The SCALE=PERCENT *plot-option* requests a plot of percentages instead of totals. The GROUPBY=ROW *plot-option* groups the graph cells by the row variable (SchoolType).

Figure 97.6 shows the two-way table produced for SchoolType by Response. The first variable named in the two-way table request, SchoolType, is referred to as the *row variable*, and the second variable, Response, is referred to as the *column variable*. Two-way tables display all column variable levels for each row variable level. This two-way table lists all levels of the column variable Response for each level of the row variable SchoolType, 'Middle School' and 'High School'. Also SchoolType = 'Total' shows the distribution of Response overall for both types of schools. And Response = 'Total' provides totals over all levels of response, for each type of school and overall. To suppress these totals, you can specify the NOTOTAL option.

Figure 97.6 Two-Way Table of SchoolType by Response

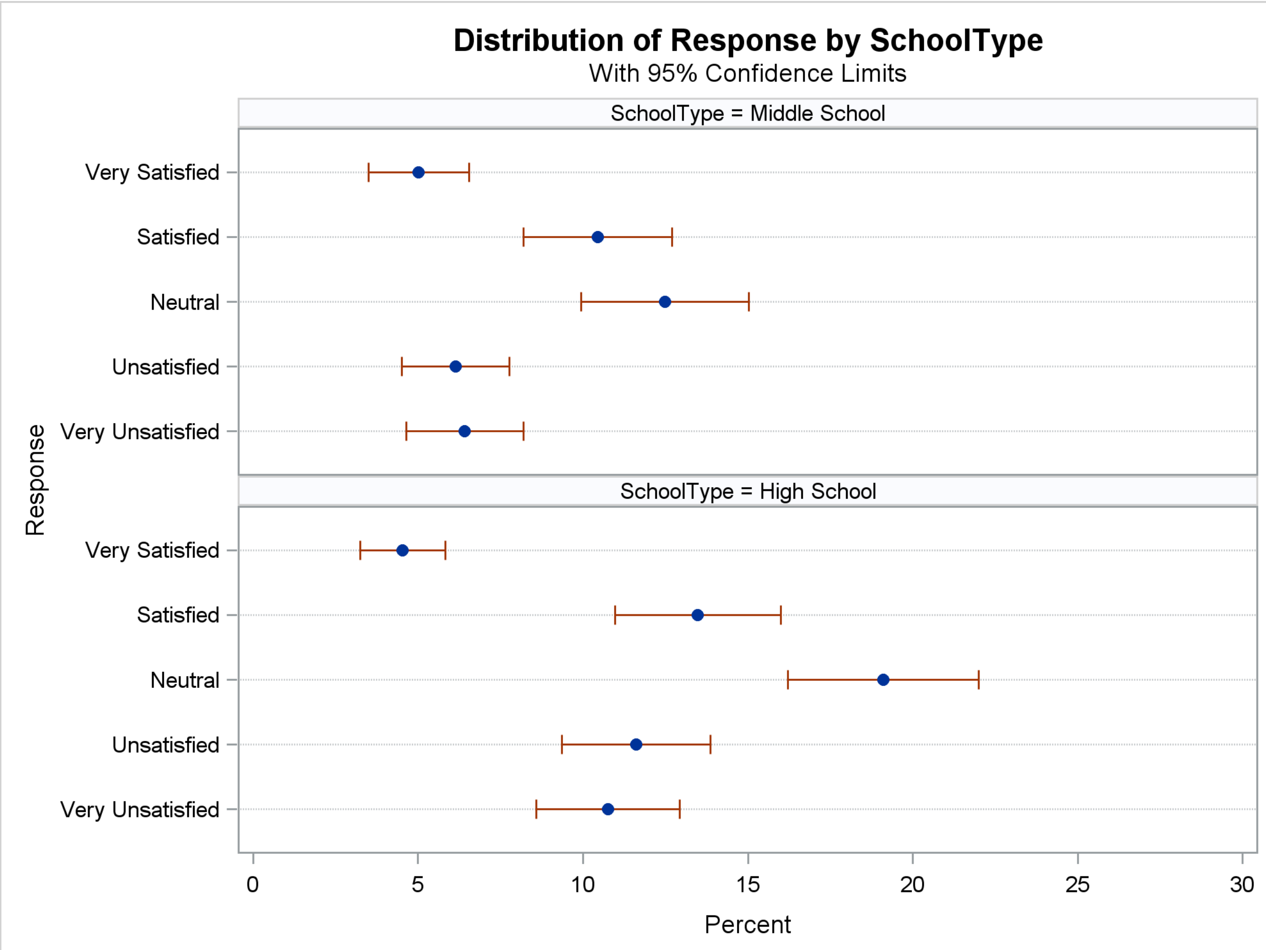
Student Information System Survey

The SURVEYFREQ Procedure

Table of SchoolType by Response						
SchoolType	Response	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
Middle School	Very Unsatisfied	116	2496	351.43834	6.4155	0.9030
	Unsatisfied	109	2389	321.97957	6.1427	0.8283
	Neutral	234	4856	504.20553	12.4847	1.2953
	Satisfied	197	4064	443.71188	10.4467	1.1417
	Very Satisfied	94	1952	302.17144	5.0193	0.7758
	Total	750	15758	1000	40.5089	2.5691
High School	Very Unsatisfied	188	4183	431.30589	10.7521	1.1076
	Unsatisfied	217	4518	446.31768	11.6137	1.1439
	Neutral	347	7434	574.17175	19.1119	1.4726
	Satisfied	258	5245	498.03221	13.4845	1.2823
	Very Satisfied	90	1762	255.67158	4.5290	0.6579
	Total	1100	23142	1003	59.4911	2.5691
Total	Very Unsatisfied	304	6678	501.61039	17.1676	1.2872
	Unsatisfied	326	6907	495.94101	17.7564	1.2712
	Neutral	581	12291	617.20147	31.5965	1.5795
	Satisfied	455	9309	572.27868	23.9311	1.4761
	Very Satisfied	184	3714	370.66577	9.5483	0.9523
	Total	1850	38900	129.85268	100.000	

Figure 97.7 displays the weighted frequency dot plot that PROC SURVEYFREQ produces for the table of SchoolType and Response. The GROUPBY=ROW *plot-option* groups the graph cells by the row variable (SchoolType). If you do not specify GROUPBY=ROW, the procedure groups the graph cells by the column variable by default. You can plot percentages instead of weighted frequencies by specifying the SCALE=PERCENT *plot-option*. You can use other *plot-options* to change the orientation of the plot or to request a different two-way layout.

Figure 97.7 Dot Plot of Percentages for SchoolType by Response



By default, without any other TABLES statement options, a two-way table displays the frequency, the weighted frequency and its standard deviation, and the percentage and its standard error for each table cell (combination of row and column variable levels). But there are several options available to customize your table display by adding more information or by suppressing some of the default information.

The following PROC SURVEYFREQ statements request a two-way table of SchoolType by Response that displays row percentages, and also request a chi-square test of association between the two variables:

```
title 'Student Information System Survey';
proc surveyfreq data=SIS_Survey nosummary;
   tables  SchoolType * Response / row nowt chisq;
   strata  State NewUser;
   cluster School;
   weight  SamplingWeight;
run;
```

The ROW option in the TABLES statement requests row percentages, which give the distribution of Response within each level of the row variable SchoolType. The NOWT option suppresses display of the weighted frequencies and their standard deviations. The CHISQ option requests a Rao-Scott chi-square test of association between SchoolType and Response.

Figure 97.8 displays the two-way table of SchoolType by Response. For middle schools, it is estimated that 25.79% of school personnel are satisfied with the student information system and 12.39% are very satisfied. For high schools, these estimates are 22.67% and 7.61%, respectively.

Figure 97.9 displays the chi-square test results. The Rao-Scott chi-square statistic equals 9.04, and the corresponding F value is 2.26 with a p-value of 0.0605. This indicates an association between school type (middle school or high school) and satisfaction with the student information system at the 10% significance level.

Figure 97.8 Two-Way Table with Row Percentages

Student Information System Survey

The SURVEYFREQ Procedure

Table of SchoolType by Response

SchoolType	Response	Frequency	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
Middle School	Very Unsatisfied	116	6.4155	0.9030	15.8373	1.9920
	Unsatisfied	109	6.1427	0.8283	15.1638	1.8140
	Neutral	234	12.4847	1.2953	30.8196	2.5173
	Satisfied	197	10.4467	1.1417	25.7886	2.2947
	Very Satisfied	94	5.0193	0.7758	12.3907	1.7449
	Total	750	40.5089	2.5691	100.000	
High School	Very Unsatisfied	188	10.7521	1.1076	18.0735	1.6881
	Unsatisfied	217	11.6137	1.1439	19.5218	1.7280
	Neutral	347	19.1119	1.4726	32.1255	2.0490
	Satisfied	258	13.4845	1.2823	22.6663	1.9240
	Very Satisfied	90	4.5290	0.6579	7.6128	1.0557
	Total	1100	59.4911	2.5691	100.000	
Total	Very Unsatisfied	304	17.1676	1.2872		
	Unsatisfied	326	17.7564	1.2712		
	Neutral	581	31.5965	1.5795		
	Satisfied	455	23.9311	1.4761		
	Very Satisfied	184	9.5483	0.9523		
	Total	1850	100.000			

Figure 97.9 Chi-Square Test of No Association

Rao-Scott Chi-Square Test	
Pearson Chi-Square	18.7829
Design Correction	2.0766
Rao-Scott Chi-Square	9.0450
DF	4
Pr > ChiSq	0.0600
F Value	2.2613
Num DF	4
Den DF	1456
Pr > F	0.0605
Sample Size = 1850	

Syntax: SURVEYFREQ Procedure

The following statements are available in the SURVEYFREQ procedure:

PROC SURVEYFREQ < *options* > ;
 BY *variables* ;
 CLUSTER *variables* ;
 REPWEIGHTS *variables* < / *options* > ;
 STRATA *variables* < / *option* > ;
 TABLES *requests* < / *options* > ;
 WEIGHT *variable* ;

The PROC SURVEYFREQ statement invokes the procedure, identifies the data set to be analyzed, and specifies the variance estimation method to use. The PROC SURVEYFREQ statement is required.

The TABLES statement specifies frequency or crosstabulation tables and requests tests and statistics for those tables. The STRATA statement lists the variables that form the strata in a stratified sample design. The CLUSTER statement specifies cluster identification variables in a clustered sample design. The WEIGHT statement names the sampling weight variable. The REPWEIGHTS statement names replicate weight variables for BRR or jackknife variance estimation. The BY statement requests completely separate analyses of groups defined by the BY variables.

All statements can appear multiple times except the PROC SURVEYFREQ statement and the WEIGHT statement, which can appear only once.

The rest of this section gives detailed syntax information for the BY, CLUSTER, REPWEIGHTS, STRATA, TABLES, and WEIGHT statements in alphabetical order after the description of the PROC SURVEYFREQ statement.

PROC SURVEYFREQ Statement

PROC SURVEYFREQ < *options* > **;**

The PROC SURVEYFREQ statement invokes the SURVEYFREQ procedure. It also identifies the data set to be analyzed, specifies the variance estimation method to use, and provides sample design information. The DATA= option names the input data set to be analyzed. The VARMETHOD= option specifies the variance estimation method, which is the Taylor series method by default. For Taylor series variance estimation, you can include a finite population correction factor in the analysis by providing either the sampling rate or population total in the RATE= or TOTAL= option, respectively. If your design is stratified with different sampling rates or totals for different strata, you can input these stratum rates or totals in a SAS data set that contains the stratification variables.

Table 97.1 summarizes the *options* available in the PROC SURVEYFREQ statement.

Table 97.1 PROC SURVEYFREQ Statement Options

Option	Description
DATA=	Names the input SAS data set
MISSING	Treats missing values as a valid level
NOMCAR	Treats missing values as *not missing completely at random*
NOSUMMARY	Suppresses the display of the "Data Summary" table
ORDER=	Specifies the order of variable levels
PAGE	Displays only one table per page
RATE=	Specifies the first-stage sampling rate
TOTAL=	Specifies the total number of primary sampling units
VARHEADER=	Specifies the variable identification to display
VARMETHOD=	Specifies the variance estimation method

You can specify the following *options* in the PROC SURVEYFREQ statement:

DATA=*SAS-data-set*

names the *SAS-data-set* to be analyzed by PROC SURVEYFREQ. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

MISSING

treats missing values as a valid (nonmissing) category for all categorical variables, which include TABLES, STRATA, and CLUSTER variables.

By default, if you do not specify the MISSING option, an observation is excluded from the analysis if it has a missing value for any STRATA or CLUSTER variable. Additionally, PROC SURVEYFREQ excludes an observation from a frequency or crosstabulation table if that observation has a missing value for any of the variables in the table request, unless you specify the MISSING option. For more information, see the section "Missing Values" on page 8001.

NOMCAR

includes observations with missing values of TABLES variables in the variance computation as *not missing completely at random* (NOMCAR) for Taylor series variance estimation. When you specify the NOMCAR option, PROC SURVEYFREQ computes variance estimates by analyzing the nonmissing values as a domain (subpopulation), where the entire population includes both nonmissing and missing domains. For more information, see the section “Missing Values” on page 8001.

By default, PROC SURVEYFREQ completely excludes an observation from a frequency or crosstabulation table (and the corresponding variance computations) if that observation has a missing value for any of the variables in the table request, unless you specify the MISSING option. The NOMCAR option has no effect when you specify the MISSING option, which treats missing values as a valid nonmissing level.

The NOMCAR option applies only to Taylor series variance estimation. The replication methods, which you can request by specifying the VARMETHOD=BRR and VARMETHOD=JACKKNIFE options, do not use the NOMCAR option.

NOSUMMARY

suppresses the display of the “Data Summary” table, which PROC SURVEYFREQ produces by default. For information about this table, see the section “Data Summary Table” on page 8038.

ORDER=DATA | FORMATTED | FREQ | INTERNAL

specifies the order of the variable levels in the frequency and crosstabulation tables, which you request in the TABLES statement. The ORDER= option also controls the order of the STRATA variable levels in the “Stratum Information” table.

The ORDER= option can take the following values:

ORDER=	Levels Ordered By
DATA	Order of appearance in the input data set
FORMATTED	External formatted value, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) value
FREQ	Descending frequency count; levels with the most observations come first in the order
INTERNAL	Unformatted value

By default, ORDER=INTERNAL. The FORMATTED and INTERNAL orders are machine-dependent. The frequency count used by ORDER=FREQ is the nonweighted frequency (sample size), rather than the weighted frequency.

For more information about sort order, see the chapter on the SORT procedure in the *Base SAS Procedures Guide* and the discussion of BY-group processing in *SAS Language Reference: Concepts*.

PAGE

displays only one table per page. Otherwise, PROC SURVEYFREQ displays multiple tables per page as space permits.

RATE=*value* | *SAS-data-set*

R=*value* | *SAS-data-set*

specifies the sampling rate, which PROC SURVEYFREQ uses to compute a finite population correction for Taylor series variance estimation. You can provide a single sampling rate *value*, or you can provide stratum sampling rates by specifying a *SAS-data-set*.

If your sample design has multiple stages, you should specify the *first-stage sampling rate*, which is the ratio of the number of primary sampling units (PSUs) in the sample to the total number of PSUs in the population.

For a nonstratified sample design, or for a stratified sample design that uses the same sampling rate in all strata, you should specify a single sampling rate *value*. If your design is stratified and uses different sampling rates in different strata, you should name a *SAS-data-set* that contains the stratification variables and the stratum sampling rates. You should provide the stratum sampling rates in the data set variable named _RATE_. For more information, see the section "Population Totals and Sampling Rates" on page 8000.

The sampling rate values must be nonnegative numbers. You can specify sampling rates as numbers between 0 and 1. Or you can specify sampling rates in percentage form as numbers between 1 and 100, which PROC SURVEYFREQ converts to proportions. The procedure treats the value 1 as 100% instead of 1%.

If you do not specify the RATE= or the TOTAL= option, the Taylor series variance estimation does not include a finite population correction. You cannot specify both the RATE= and the TOTAL= option in the same PROC SURVEYFREQ statement.

PROC SURVEYSELECT does not use the RATE= or the TOTAL= option for BRR or jackknife variance estimation (which you can request by specifying the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option, respectively).

TOTAL=*value* | *SAS-data-set*

N=*value* | *SAS-data-set*

specifies the total number of primary sampling units (PSUs), which PROC SURVEYFREQ uses to compute a finite population correction for Taylor series variance estimation. You can provide a single total *value*, or you can provide stratum totals by specifying a *SAS-data-set*. The totals must be positive numbers.

If your sample design has multiple stages, you should specify the total number of primary sampling units (PSUs).

For a nonstratified sample design, you should specify a single total *value*, which refers to the total number of PSUs in the population. For a stratified sample design that has the same population total in each stratum, you can specify a single total *value*, which refers to the total number of PSUs in each stratum. If your design is stratified and has different totals in different strata, you should name a *SAS-data-set* that contains the stratification variables and the stratum totals. You should provide the stratum totals in the data set variable named _TOTAL_. For more information, see the section "Population Totals and Sampling Rates" on page 8000.

If you do not specify the RATE= or the TOTAL= option, the Taylor series variance estimation does not include a finite population correction. You cannot specify both the RATE= and the TOTAL= option in the same PROC SURVEYFREQ statement.

PROC SURVEYSELECT does not use the RATE= or the TOTAL= option for BRR or jackknife variance estimation (which you can request by specifying the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option, respectively).

VARHEADER=LABEL | NAME | NAMELABEL

specifies the variable identification to use in the displayed output. By default VARHEADER=NAME, which displays variable names in the output. The VARHEADER= option affects the headers of the variable level columns in one-way frequency tables, crosstabulation tables, and the “Stratum Information” table. The VARHEADER= option also controls variable identification in the table headers.

The VARHEADER= option can take the following values:

VARHEADER=	Variable Identification Displayed
LABEL	Variable label
NAME	Variable name
NAMELABEL	Variable name and label, as *Name* (*Label*)

VARMETHOD=BRR < (*method-options*) >

VARMETHOD=JACKKNIFE | JK < (*method-options*) >

VARMETHOD=TAYLOR

specifies the variance estimation method. VARMETHOD=TAYLOR requests the Taylor series method, which is the default if you do not specify the VARMETHOD= option or the REPWEIGHTS statement. VARMETHOD=BRR requests variance estimation by balanced repeated replication (BRR), and VARMETHOD=JACKKNIFE requests variance estimation by the delete-1 jackknife method.

For VARMETHOD=BRR and VARMETHOD=JACKKNIFE, you can specify *method-options* in parentheses after the variance method name. For example:

```
varmethod=BRR(reps=60 outweights=myReplicateWeights)
```

Table 97.2 summarizes the available *method-options*.

Table 97.2 Variance Estimation Options

VARMETHOD=	Variance Estimation Method	Method Options
BRR	Balanced repeated replication	DFADJ FAY <=*value*> HADAMARD=*SAS-data-set* OUTWEIGHTS=*SAS-data-set* PRINTH REPS=*number*
JACKKNIFE | JK	Jackknife	DFADJ OUTJKCOEFS=*SAS-data-set* OUTWEIGHTS=*SAS-data-set*
TAYLOR	Taylor series linearization	None

You can specify the following values for the VARMETHOD= option:

BRR < (*method-options*) >

requests variance estimation by balanced repeated replication (BRR). The BRR method requires a stratified sample design that has two primary sampling units (PSUs) in each stratum. If you specify this option, you must also specify a STRATA statement unless you use a REPWEIGHTS statement to provide replicate weights. For more information, see the section “Balanced Repeated Replication (BRR)” on page 8011.

You can specify the following *method-options*:

DFADJ

computes the degrees of freedom as the number of nonmissing strata for the individual table request. If you specify this option, PROC SURVEYFREQ does not count any empty strata that occur when observations that have missing values of the TABLES variables are removed from the analysis of the table. By default, PROC SURVEYFREQ computes the degrees of freedom by counting the number of nonmissing strata for all valid observations in the input data set.

For more information, see the section “Degrees of Freedom” on page 8019. For information about valid observations, see the section “Data Summary Table” on page 8038.

This *method-option* has no effect when you specify the MISSING option, which treats missing values as a valid nonmissing level.

This *method-option* is not used when you specify the degrees of freedom in the DF= option in the TABLES statement or when you specify a REPWEIGHTS statement to provide replicate weights. When you specify a REPWEIGHTS statement, the degrees of freedom are the number of REPWEIGHTS variables (replicates) unless you specify the DF= option in the REPWEIGHTS or the TABLES statement.

FAY < =*value* >

requests Fay’s method, which is a modification of the BRR method. For more information, see the section “Fay’s BRR Method” on page 8012.

You can specify the *value* of the Fay coefficient, which is used in converting the original sampling weights to replicate weights. The Fay coefficient must be a nonnegative number less than 1. By default, the Fay coefficient is 0.5.

HADAMARD=*SAS-data-set*

H=*SAS-data-set*

names a *SAS-data-set* that contains the Hadamard matrix for BRR replicate construction. If you do not specify this *method-option*, PROC SURVEYFREQ generates an appropriate Hadamard matrix for replicate construction. For more information, see the sections “Balanced Repeated Replication (BRR)” on page 8011 and “Hadamard Matrix” on page 8013.

If a Hadamard matrix of a particular dimension exists, it is not necessarily unique. Therefore, if you want to use a specific Hadamard matrix, you must provide the matrix as a *SAS-data-set* in this *method-option*.

In the HADAMARD= input data set, each variable corresponds to a column and each observation corresponds to a row of the Hadamard matrix. You can use any variable names in the HADAMARD= data set. All values in the data set must equal either 1 or –1. You must ensure that the matrix you provide is indeed a Hadamard matrix—that is, $\mathbf{A}'\mathbf{A} = R\mathbf{I}$, where $\mathbf{A}$ is the Hadamard matrix of dimension R and $\mathbf{I}$ is an identity matrix. PROC SURVEYFREQ does not check the validity of the Hadamard matrix that you provide.

The HADAMARD= input data set must contain at least H variables, where H denotes the number of first-stage strata in your design. If the data set contains more than H variables, PROC SURVEYFREQ uses only the first H variables. Similarly, the HADAMARD= input data set must contain at least H observations.

If you do not specify the REPS= *method-option*, the number of replicates is assumed to be the number of observations in the HADAMARD= input data set. If you specify the number of replicates—for example, REPS=*nreps*—the first *nreps* observations in the HADAMARD= data set are used to construct the replicates.

You can specify the PRINTH *method-option* to display the Hadamard matrix that PROC SURVEYFREQ uses to construct replicates for BRR.

OUTWEIGHTS=*SAS-data-set*

names a *SAS-data-set* to store the replicate weights that PROC SURVEYFREQ creates for BRR variance estimation. For information about replicate weights, see the section "Balanced Repeated Replication (BRR)" on page 8011. For information about the contents of the OUTWEIGHTS= data set, see the section "Replicate Weight Output Data Set" on page 8037.

The OUTWEIGHTS= *method-option* is not available when you provide replicate weights in a REPWEIGHTS statement.

PRINTH

displays the Hadamard matrix that PROC SURVEYFREQ uses to construct replicates for BRR variance estimation. When you provide the Hadamard matrix in the HADAMARD= *method-option*, PROC SURVEYFREQ displays only the rows and columns that are actually used to construct replicates. For more information, see the sections "Balanced Repeated Replication (BRR)" on page 8011 and "Hadamard Matrix" on page 8013.

The PRINTH *method-option* is not available when you provide replicate weights in a REPWEIGHTS statement because the procedure does not use a Hadamard matrix in this case.

REPS=*number*

specifies the *number* of replicates for BRR variance estimation. The value of *number* must be an integer greater than 1.

If you do not use the HADAMARD= *method-option* to provide a Hadamard matrix, the number of replicates should be greater than the number of strata and should be a multiple of 4. For more information, see the section "Balanced Repeated Replication (BRR)" on page 8011. If PROC SURVEYFREQ cannot construct a Hadamard matrix for the REPS= value that you specify, the value is increased until a Hadamard matrix of that dimension can be constructed. Therefore, the actual number of replicates that PROC SURVEYFREQ uses might be larger than *number*.

If you use the HADAMARD= *method-option* to provide a Hadamard matrix, the value of *number* must not be less than the number of rows in the Hadamard matrix. If you provide a Hadamard matrix and do not specify the REPS= *method-option*, the number of replicates equals the number of rows in the Hadamard matrix.

If you do not specify the REPS= or the HADAMARD= *method-option* and do not use a REPWEIGHTS statement, the number of replicates equals the smallest multiple of 4 that is greater than the number of strata.

If you use a REPWEIGHTS statement to provide replicate weights, PROC SURVEYFREQ does not use the REPS= *method-option*; the number of replicates equals the number of REPWEIGHTS variables.

JACKKNIFE < (*method-options*) >

JK < (*method-options*) >

requests variance estimation by the delete-1 jackknife method. For more information, see the section "The Jackknife Method" on page 8014. If you use a REPWEIGHTS statement to provide replicate weights, VARMETHOD=JACKKNIFE is the default variance estimation method.

The delete-1 jackknife method requires at least two primary sampling units (PSUs) in each stratum for stratified designs unless you use a REPWEIGHTS statement to provide replicate weights.

You can specify the following *method-options*:

DFADJ

computes the degrees of freedom by using the number of nonmissing strata and clusters for the individual table request. If you specify this *method-option*, PROC SURVEYFREQ does not count any empty strata or clusters that occur when observations that have missing values of the TABLES variables are removed from the analysis of the table. By default, PROC SURVEYFREQ computes the degrees of freedom by counting the number of nonmissing strata and clusters for all valid observations in the input data set. The degrees of freedom for VARMETHOD=JACKKNIFE equal the number of clusters minus the number of strata.

For more information, see the section "Degrees of Freedom" on page 8019. For information about valid observations, see the section "Data Summary Table" on page 8038.

This *method-option* has no effect when you specify the MISSING option, which treats missing values as a valid nonmissing level.

This *method-option* is not used when you specify the degrees of freedom in the DF= option in the TABLES statement or when you specify a REPWEIGHTS statement to provide replicate weights. When you specify a REPWEIGHTS statement, the degrees of freedom are the number of REPWEIGHTS variables (replicates) unless you specify the DF= option in the REPWEIGHTS or the TABLES statement.

OUTJKCOEFS=*SAS-data-set*

names a *SAS-data-set* to store the jackknife coefficients. For information about jackknife coefficients, see the section "The Jackknife Method" on page 8014. For information about the contents of the OUTJKCOEFS= data set, see the section "Jackknife Coefficient Output Data Set" on page 8038.

OUTWEIGHTS=*SAS-data-set*

names a *SAS-data-set* to store the replicate weights that PROC SURVEYFREQ creates for jackknife variance estimation. For information about replicate weights, see the section "The Jackknife Method" on page 8014. For information about the contents of the OUTWEIGHTS= data set, see the section "Replicate Weight Output Data Set" on page 8037.

This *method-option* is not available when you use a REPWEIGHTS statement to provide replicate weights.

TAYLOR

requests Taylor series variance estimation. This is the default method if you do not specify the VARMETHOD= option or a REPWEIGHTS statement. For more information, see the section "Taylor Series Variance Estimation" on page 8004.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC SURVEYFREQ to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement for the SURVEYFREQ procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

Using a BY statement provides completely separate analyses of the BY groups. It does not provide a statistically valid domain (subpopulation) analysis, where the total number of units in the subpopulation is not known with certainty. You should include the domain variable(s) in your TABLES request to obtain domain analysis. For more information, see the section "Domain Analysis" on page 8001.

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

CLUSTER Statement

CLUSTER *variables* **;**

The CLUSTER statement names one or more *variables* that identify the first-stage clusters in a clustered sample design. First-stage clusters are also known as primary sampling units (PSUs). The combinations of levels of the CLUSTER variables define the clusters in the sample. If there is a STRATA statement, clusters are nested within strata.

If your sample design has clustering at multiple stages, you should specify only the first-stage clusters (PSUs) in the CLUSTER statement. See the section "Specifying the Sample Design" on page 7999 for more information.

If you provide replicate weights for BRR or jackknife variance estimation by using the REPWEIGHTS statement, you do not need to specify a CLUSTER statement.

The CLUSTER variables are one or more variables in the DATA= input data set. These variables can be either character or numeric, but the procedure treats them as categorical variables. The formatted values of the CLUSTER variables determine the CLUSTER variable levels. Thus, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the *Base SAS Procedures Guide* and the discussions of the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference*.

An observation is excluded from the analysis if it has a missing value for any CLUSTER variable unless you specify the MISSING option in the PROC SURVEYFREQ statement. For more information, see the section "Missing Values" on page 8001.

You can use multiple CLUSTER statements to specify CLUSTER variables. The procedure uses variables from all CLUSTER statements to create clusters.

REPWEIGHTS Statement

REPWEIGHTS *variables* < */ options* > **;**

The REPWEIGHTS statement names *variables* that provide replicate weights for BRR or jackknife variance estimation, which you can request by specifying the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option in the PROC SURVEYFREQ statement. If you do not provide replicate weights for these methods by using a REPWEIGHTS statement, then PROC SURVEYFREQ constructs replicate weights for the analysis. See the sections "Balanced Repeated Replication (BRR)" on page 8011 and "The Jackknife Method" on page 8014 for information about replicate weights.

Each REPWEIGHTS variable should contain the weights for a single replicate, and the number of replicates equals the number of REPWEIGHTS variables. The REPWEIGHTS variables must be numeric, and the variable values must be nonnegative numbers.

If you provide replicate weights by using a REPWEIGHTS statement, you do not need to specify a CLUSTER or STRATA statement. If you use a REPWEIGHTS statement and do not specify the VARMETHOD= option in the PROC SURVEYFREQ statement, the procedure uses VARMETHOD=JACKKNIFE by default.

If you specify a REPWEIGHTS statement but do not include a WEIGHT statement, PROC SURVEYFREQ uses the average of each observation's replicate weights as the observation's weight.

You can specify the following *options* in the REPWEIGHTS statement after a slash (/):

DF=*df*

specifies the degrees of freedom for the analysis. The value of *df* must be a positive number. By default, the degrees of freedom equal the number of REPWEIGHTS variables. For more information, see the section "Degrees of Freedom" on page 8019.

PROC SURVEYFREQ uses the value *df* to obtain the *t*-percentile for confidence limits for proportions, totals, and other statistics. For more information, see the section "Confidence Limits for Proportions" on page 8016. PROC SURVEYFREQ also uses *df* to compute the denominator degrees of freedom for the *F* statistics in the Rao-Scott and Wald chi-square tests. For more information, see the sections "Rao-Scott Chi-Square Test" on page 8028, "Rao-Scott Likelihood Ratio Chi-Square Test" on page 8033, "Wald Chi-Square Test" on page 8035, and "Wald Log-Linear Chi-Square Test" on page 8036.

JKCOEFS=*value* **|** **<** **(** **>***values* **<** **)** **>** **|** *SAS-data-set*

specifies the jackknife coefficients for jackknife variance estimation (which you can request by specifying VARMETHOD=JACKKNIFE). You can provide a single jackknife coefficient *value* to use for all replicates, or you can provide a value for each replicate by specifying a list of *values* or a *SAS-data-set*. The jackknife coefficient values must be nonnegative numbers. For more information, see the section "The Jackknife Method" on page 8014.

You can provide jackknife coefficients by specifying one of the following forms:

value

specifies a single jackknife coefficient *value* to use for all replicates. The coefficient value must be a nonnegative number.

values

specifies a list of jackknife coefficient *values*, where each value corresponds to a single replicate that is identified by a REPWEIGHTS variable. You can separate the values with blanks or commas, and you can enclose the list of values in parentheses. The coefficient values must be nonnegative numbers. The number of coefficient values should equal the number of replicate weight variables that you specify in the REPWEIGHTS statement.

You should list the jackknife coefficient values in the same order in which you list the corresponding replicate weight variables in the REPWEIGHTS statement.

SAS-data-set

names a *SAS-data-set* that contains the jackknife coefficients. You should provide the jackknife coefficients in the data set variable named JKCoefficient. Each coefficient value must be a nonnegative number. Each observation in this data set should correspond to a replicate that is identified by a REPWEIGHTS variable. The number of observations in this data set must not be less than the number of REPWEIGHTS variables.

STRATA Statement

STRATA *variables* < */ option* > **;**

The STRATA statement names one or more *variables* that identify the first-stage strata in a stratified sample design. The combinations of levels of STRATA variables define the strata in the sample, where strata are nonoverlapping subgroups that were sampled independently.

If your sample design has stratification at multiple stages, you should specify only the first-stage strata in the STRATA statement. For more information, see the section "Specifying the Sample Design" on page 7999.

If you use a REPWEIGHTS statement to provide replicate weights for BRR or jackknife variance estimation, you do not need to specify a STRATA statement.

The STRATA variables are one or more variables in the DATA= input data set. These variables can be either character or numeric, but the procedure treats them as categorical variables. The formatted values of the STRATA variables determine the STRATA variable levels. Thus, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the *Base SAS Procedures Guide* and the discussions of the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference*.

PROC SURVEYFREQ excludes an observation from the analysis if it has a missing value for any STRATA variable unless you specify the MISSING option in the PROC SURVEYFREQ statement. For more information, see the section "Missing Values" on page 8001.

You can use multiple STRATA statements to specify STRATA variables. The procedure uses variables from all STRATA statements to define strata.

You can specify the following *option* in the STRATA statement after a slash (/):

LIST
displays the "Stratum Information" table, which lists all strata together with the corresponding values of the STRATA variables. This table provides the number of observations and the number of clusters in each stratum, as well as the sampling fraction if you specify the RATE= or TOTAL= option in the PROC SURVEYFREQ statement. For more information, see the section "Stratum Information Table" on page 8039.

TABLES Statement

TABLES *requests* < */ options* > **;**

The TABLES statement requests one-way to *n*-way frequency and crosstabulation tables and statistics for these tables.

If you omit the TABLES statement, PROC SURVEYFREQ generates one-way frequency tables for all DATA= data set variables that are not listed in the other statements.

The following argument is required in the TABLES statement:

requests
specify the frequency and crosstabulation tables to produce. A *request* is composed of one variable name or several variable names separated by asterisks. To request a one-way frequency table, use a single variable. To request a two-way crosstabulation table, use an asterisk between two variables. To request a multiway table (an n-way table, where $n > 2$), separate the desired variables with asterisks. The unique values of these variables form the rows, columns, and layers of the table.

For two-way tables to multiway tables, the values of the last variable form the crosstabulation table columns, while the values of the next-to-last variable form the rows. Each level (or combination of levels) of the other variables forms one layer. PROC SURVEYFREQ produces a separate crosstabulation table for each layer. For example, a specification of A*B*C*D in a TABLES statement produces k tables, where k is the number of different combinations of levels for A and B. Each table lists the levels for D (columns) within each level of C (rows).

You can use multiple TABLES statements in a single PROC SURVEYFREQ step. You can also specify any number of table requests in a single TABLES statement. To specify multiple table requests quickly, use a grouping syntax by placing parentheses around several variables and joining other variables or variable combinations. Table 97.3 shows some examples of grouping syntax.

Table 97.3 Grouping Syntax

TABLES Request	Equivalent to
A*(B C)	A*B A*C
(A B)*(C D)	A*C B*C A*D B*D
(A B C)*D	A*D B*D C*D
A – – C	A B C
(A – – C)*D	A*D B*D C*D

The TABLES statement variables are one or more variables from the DATA= input data set. These variables can be either character or numeric, but the procedure treats them as categorical variables. PROC SURVEYFREQ uses the formatted values of the TABLES variable to determine the categorical variable levels. If you assign a format to a variable by using a FORMAT statement, PROC SURVEYFREQ formats the values before dividing observations into the levels of a frequency or crosstabulation table. See the discussion of the FORMAT procedure in the *Base SAS Procedures Guide* and the discussions of the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference*.

By default, the frequency or crosstabulation table lists the values of both character and numeric variables in ascending order based on internal (unformatted) variable values. You can change the order of the values in the table by specifying the ORDER= option in the PROC SURVEYFREQ statement. To list the values in ascending order by formatted value, use ORDER=FORMATTED.

Without Options

If you request a frequency or crosstabulation table without specifying options, PROC SURVEYFREQ produces the following for each table level or cell:

- frequency (sample size)
- weighted frequency, which estimates the population total
- standard deviation of the weighted frequency
- percentage, which estimates the population proportion
- standard error of the percentage

The table displays weighted frequencies if your analysis includes a WEIGHT statement, or if you specify the WTFREQ option in the TABLES statement. The table also displays the number of observations that have missing values. For more information, see the sections “One-Way Frequency Tables” on page 8040 and “Crosstabulation Tables” on page 8041.

Options

Table 97.4 summarizes the *options* available in the TABLES statement. Descriptions of the *options* follow the table in alphabetical order.

Table 97.4 TABLES Statement Options

Option	Description
Control Statistical Analysis	
AGREE	Requests kappa coefficients
ALPHA=	Sets level for confidence limits
CHISQ	Requests Rao-Scott chi-square test
CL	Requests confidence limits for percentages and specifies confidence limit type for percentages
CLWT	Requests confidence limits for weighted frequencies
COV	Requests covariances of frequency estimates
COVP	Requests covariances of proportion estimates
DF=	Specifies degrees of freedom
KAPPA	Requests simple kappa coefficient
LRCHISQ	Requests Rao-Scott likelihood ratio test
OR	Requests odds ratio and relative risks
RISK	Requests risks and risk difference
TESTP=	Specifies null proportions for one-way chi-square test
WCHISQ	Requests Wald chi-square test
WLLCHISQ	Requests Wald log-linear chi-square test
WTKAPPA	Requests weighted kappa coefficient
Request Additional Table Information	
CELLCHI2	Displays cell contributions to the Pearson chi-square
CLWT	Displays confidence limits for weighted frequencies
COLUMN	Displays column percentages and standard errors
CV	Displays coefficients of variation for percentages
CVWT	Displays coefficients of variation for weighted frequencies
DEFF	Displays design effects for percentages
DEVIATION	Displays deviations of weighted frequencies
EXPECTED	Displays expected weighted frequencies
PEARSONRES	Displays Pearson residuals
ROW	Displays row percentages and standard errors
VAR	Displays variances of percentages
VARWT	Displays variances of weighted frequencies
WTFREQ	Displays totals and standard errors when there is no WEIGHT statement
Control Displayed Output	
NOCELLPERCENT	Suppresses display of overall percentages
NOFREQ	Suppresses display of frequency counts
NOPERCENT	Suppresses display of all percentages
NOPRINT	Suppresses display of tables but displays statistical tests
NOSPARSE	Suppresses display of zero rows and columns
NOSTD	Suppresses display of standard errors for all estimates
NOTOTAL	Suppresses display of row and column totals
NOWT	Suppresses display of weighted frequencies
Produce Statistical Graphics	
PLOTS=	Requests plots from ODS Graphics

You can specify the following *options* in a TABLES statement:

AGREE < (*options*) >

requests the simple and weighted kappa coefficients with their standard errors and confidence limits. Kappa coefficients can be computed for square two-way tables, where the number of rows equals the number of columns. For 2×2 tables, the weighted kappa coefficient equals the simple kappa coefficient, and PROC SURVEYFREQ displays only the simple kappa coefficient. For more information, see the section "Kappa Coefficients" on page 8025.

Kappa coefficients are available when you specify variance estimation by the jackknife method (VARMETHOD=JACKKNIFE) or by balanced repeated replication (VARMETHOD=BRR); kappa coefficients are not available with the Taylor series method (VARMETHOD=TAYLOR).

The weighted kappa coefficient is computed by using agreement weights that reflect the relative agreement between pairs of variable levels. Agreement weights are not the same as *sampling weights*, which you provide by specifying the WEIGHT statement. PROC SURVEYFREQ uses sampling weights to compute both the simple and weighted kappa coefficients. For more information, see the section "Weighted Kappa Coefficient" on page 8026.

You can specify the level for the kappa confidence limits in the ALPHA= option. By default, ALPHA=0.05, which produces 95% confidence limits.

You can request the simple kappa coefficient or the weighted kappa coefficient separately by specifying the KAPPA or WTKAPPA option, respectively.

You can specify the following *options*:

PRINTKWTS

displays the agreement weights that PROC SURVEYFREQ uses to compute the weighted kappa coefficient. Agreement weights reflect the relative agreement between pairs of variable levels. By default, PROC SURVEYFREQ uses the Cicchetti-Allison form of agreement weights. If you specify the WT=FC option, the procedure uses the Fleiss-Cohen form of agreement weights. For more information, see the section "Weighted Kappa Coefficient" on page 8026.

WT=FC

requests Fleiss-Cohen agreement weights for the weighted kappa computation. By default, PROC SURVEYFREQ uses Cicchetti-Allison agreement weights to compute the weighted kappa coefficient. Agreement weights reflect the relative agreement between pairs of variable levels. For more information, see the section "Weighted Kappa Coefficient" on page 8026.

ALPHA=α

specifies the level for confidence limits. The value of α must be between 0 and 1; a confidence level of α produces $100(1-\alpha)\%$ confidence limits. By default, ALPHA=0.05, which produces 95% confidence limits.

You can request confidence limits for percentages by specifying the CL option, and you can request confidence limits for weighted frequencies by specifying the CLWT option. For more information, see the sections "Confidence Limits for Proportions" on page 8016 and "Confidence Limits for Totals" on page 8016.

The ALPHA= option also applies to confidence limits for the risks and risk difference (which you can request by specifying the RISK option) and to confidence limits for the odds ratio and relative risks (which you can request by specifying the OR option). For more information, see the sections "Risks and Risk Difference" on page 8022 and "Odds Ratio and Relative Risks" on page 8023.

CELLCHI2

displays each table cell's contribution to the Pearson chi-square statistic in the crosstabulation table. The cell chi-square is computed as $(\textit{weighted frequency} - \textit{expected})^2 \; / \; \textit{expected}$, where *weighted frequency* is the weighted frequency of the table cell and *expected* is the expected weighted frequency, which is computed under the null hypothesis that the row and column variables are independent. You can display the expected weighted frequencies by specifying the EXPECTED option, and you can display the deviations (*weighted frequency – expected*) by specifying the DEVIATION option. For more information, see the sections "Expected Weighted Frequency" on page 8021 and "Rao-Scott Chi-Square Test" on page 8028. This option has no effect for one-way tables.

CHISQ < (*options*) >

requests the Rao-Scott chi-square test. This is a design-adjusted test that is computed by applying a design correction to the weighted Pearson chi-square statistic. By default, PROC SURVEYFREQ provides a first-order Rao-Scott chi-square test. If you specify CHISQ(SECONDORDER), the procedure provides a second-order (Satterthwaite) Rao-Scott chi-square test. For more information, see the section "Rao-Scott Chi-Square Test" on page 8028.

For one-way tables, the CHISQ option produces a design-based goodness-of-fit test. By default, this is a goodness-of-fit test for equal proportions. If you specify the null hypothesis proportions in the TESTP= option, the CHISQ option produces a chi-square goodness-of-fit test for the specified proportions.

By default for one-way tables, and for first-order tests for two-way tables, the design correction is computed from proportion estimates. If you specify CHISQ(MODIFIED), the design correction is computed from null hypothesis proportions. For second-order tests for two-way tables, the design correction is always computed from null hypothesis proportions.

You can specify the following *options*:

FIRSTORDER

requests a first-order Rao-Scott chi-square test. This is the default for the CHISQ option; if you do not specify CHISQ(SECONDORDER), the procedure provides a first-order Rao-Scott test.

MODIFIED

uses the null hypothesis proportions to compute the Rao-Scott design correction. By default (if you do not specify CHISQ(MODIFIED)), the procedure uses proportion estimates to compute the design correction for all first-order tests and for second-order tests for one-way tables. For second-order tests for two-way tables, the procedure always uses null hypothesis proportions to compute the design correction.

SECONDORDER

requests a second-order (Satterthwaite) Rao-Scott chi-square test. For more information, see the section "Rao-Scott Chi-Square Test" on page 8028.

CL < (*options*) >

requests confidence limits for the percentages (proportions) in the crosstabulation table. By default, PROC SURVEYFREQ computes standard Wald ("linear") confidence limits for proportions by using the variance estimates that are based on the sample design. For more information, see the section "Confidence Limits for Proportions" on page 8016. You can specify the confidence level in the ALPHA= option. By default, ALPHA=0.05, which produces 95% confidence limits.

You can specify *options* in parentheses after the CL option to control the confidence limit computations. You can use the TYPE= option to request an alternative confidence limit type. In addition to Wald confidence limits, the following types of design-based confidence limits are available for proportions: modified Clopper-Pearson (exact), modified Wilson (score), and logit confidence limits.

If you specify the PSMALL option, PROC SURVEYFREQ uses the alternative confidence limit type for extreme (small or large) proportion estimates and uses Wald confidence limits for all other proportion estimates. If you do not specify the PSMALL option, PROC SURVEYFREQ computes the specified confidence limit type for all proportion values.

You can specify the following *options*:

ADJUST=NO | YES

controls the degrees-of-freedom adjustment to the effective sample size for the modified Clopper-Pearson and Wilson confidence limits. By default, ADJUST=YES. If you specify ADJUST=NO, the confidence limit computations do not apply the degrees-of-freedom adjustment to the effective sample size. For more information, see the section "Modified Confidence Limits" on page 8017.

The ADJUST= option is available for TYPE=CLOPPERPEARSON and TYPE=WILSON confidence limits.

PSMALL < =*p* >

uses the alternative confidence limit type that you specify in the TYPE= option for extreme (small or large) proportion values.

The PSMALL value *p* defines the range of extreme proportion values, where those proportions less than or equal to *p* or greater than or equal to (1 – *p*) are considered to be extreme, and those proportions between *p* and (1 – *p*) are not extreme. If you do not specify a PSMALL value *p*, PROC SURVEYFREQ uses $p = 0.25$ by default. For $p = 0.25$, the procedure computes Wald confidence limits for proportions between 0.25 and 0.75 and computes the alternative confidence limit type for proportions less than or equal to 0.25 or greater than or equal to 0.75.

The PSMALL value *p* must be a nonnegative number. You can specify *p* as a proportion between 0 and 0.5. Or you can specify *p* in percentage form as a number between 1 and 50, and PROC SURVEYFREQ converts that number to a proportion. The procedure treats the value 1 as the percentage form 1%.

The PSMALL option is available for TYPE=CLOPPERPEARSON, TYPE=LOGIT, and TYPE=WILSON confidence limits. For more information, see the section "Confidence Limits for Proportions" on page 8016.

TRUNCATE=NO | YES

controls the truncation of the effective sample size for the modified Clopper-Pearson and Wilson confidence limits. By default, TRUNCATE=YES truncates the effective sample size if it is larger than the original sample size. If you specify TRUNCATE=NO, the effective sample size is not truncated. For more information, see the section "Modified Confidence Limits" on page 8017.

The TRUNCATE= option is available for TYPE=CLOPPERPEARSON and TYPE=WILSON confidence limits.

TYPE=*type*

specifies the *type* of confidence limits to compute for proportions. If you do not specify the TYPE= option, PROC SURVEYFREQ computes Wald confidence limits (TYPE=WALD) by default.

If you specify the CL(PSMALL) option, the procedure uses the specified confidence limit type for extreme proportions (outside the PSMALL range) and uses Wald confidence limits for proportions that are not outside the range. If you do not specify the CL(PSMALL) option, the procedure uses the specified confidence limit type for all proportions.

You can specify one of the following confidence limit *types*:

CLOPPERPEARSON

CP

requests modified Clopper-Pearson (exact) confidence limits for proportions. For more information, see the section "Modified Clopper-Pearson Confidence Limits" on page 8018.

LOGIT

requests logit confidence limits for proportions. For more information, see the section "Logit Confidence Limits" on page 8018.

WALD

requests standard Wald ("linear") confidence limits for proportions. This is the default confidence limit type if you do not specify the TYPE= option. For more information, see the section "Wald Confidence Limits" on page 8017.

WILSON

SCORE

requests modified Wilson (score) confidence limits for proportions. For more information, see the section "Modified Wilson Confidence Limits" on page 8018.

CLWT

requests confidence limits for the weighted frequencies (totals) in the crosstabulation table. You can specify the confidence level in the ALPHA= option. By default, ALPHA=0.05, which produces 95% confidence limits. For more information, see the section "Confidence Limits for Totals" on page 8016.

COLUMN < (*option***) >**

displays the column percentage (estimated proportion of the column total) for each cell in a two-way table. The COLUMN option also provides the standard errors of the column percentages. For more information, see the section "Row and Column Proportions" on page 8010. This option has no effect for one-way tables.

You can specify the following *option*:

DEFF

displays the design effect for each column percentage in the crosstabulation table. For more information, see the section "Design Effect" on page 8020.

COV

requests the covariance matrix of the table cell frequency estimates. For more information, see the section "Covariances of Frequency Estimates" on page 8008.

COVP

requests the covariance matrix of the proportion estimates.

CV

displays the coefficient of variation for each percentage (proportion) estimate in the crosstabulation table. For more information, see the section "Coefficient of Variation" on page 8020.

CVWT

displays the coefficient of variation for each weighted frequency (estimated total), in the crosstabulation table. For more information, see the section "Coefficient of Variation" on page 8020.

DEFF

displays the design effect for each overall percentage (proportion) estimate in the crosstabulation table. For more information, see the section "Design Effect" on page 8020.

To request design effects for row or column percentages, specify the DEFF option in parentheses after the ROW or COLUMN option.

DEVIATION

displays the deviations of the weighted frequencies from the expected weighted frequencies (*weighted frequency – expected*) in the crosstabulation table. The expected weighted frequencies are computed under the null hypothesis that the row and column variables are independent. You can display the expected values by specifying the EXPECTED option. For more information, see the section "Expected Weighted Frequency" on page 8021. This option has no effect for one-way tables.

DF=*df*

specifies the degrees of freedom for the analysis. The value of *df* must be a nonnegative number. By default, PROC SURVEYFREQ computes the degrees of freedom as described in the section "Degrees of Freedom" on page 8019.

PROC SURVEYFREQ uses the value *df* to obtain the *t*-percentile for confidence limits for proportions, totals, and other statistics. For more information, see the section "Confidence Limits for Proportions" on page 8016. PROC SURVEYFREQ also uses *df* to compute the denominator degrees of freedom for the *F* statistics in the Rao-Scott and Wald chi-square tests. For more information, see the sections "Rao-Scott Chi-Square Test" on page 8028, "Rao-Scott Likelihood Ratio Chi-Square Test" on page 8033, "Wald Chi-Square Test" on page 8035, and "Wald Log-Linear Chi-Square Test" on page 8036.

EXPECTED

displays the expected weighted frequencies for the cells in the crosstabulation table. The expected weighted frequencies are computed under the null hypothesis that the row and column variables are independent. For more information, see the section "Expected Weighted Frequency" on page 8021. This option has no effect for one-way tables.

KAPPA

requests the simple kappa coefficient with its standard error and confidence limits. The kappa coefficient can be computed for square two-way tables, where the number of rows equals the number of columns. For more information, see the section "Simple Kappa Coefficient" on page 8025.

The kappa coefficient is available when you specify variance estimation by the jackknife method (VARMETHOD=JACKKNIFE) or by balanced repeated replication (VARMETHOD=BRR); the kappa coefficient is not available with the Taylor series method (VARMETHOD=TAYLOR).

You can specify the level for the kappa confidence limits in the ALPHA= option. By default, ALPHA=0.05, which produces 95% confidence limits.

LRCHISQ < (*options*) >

requests the Rao-Scott likelihood ratio chi-square test. This is a design-adjusted test that is computed by applying a design correction to the weighted likelihood ratio chi-square statistic. By default, PROC SURVEYFREQ provides a first-order Rao-Scott likelihood ratio test. If you specify LRCHISQ(SECONDORDER), the procedure provides a second-order (Satterthwaite) Rao-Scott likelihood ratio test. For more information, see the section “Rao-Scott Likelihood Ratio Chi-Square Test” on page 8033.

For one-way tables, the LRCHISQ option produces a design-based likelihood ratio goodness-of-fit test. By default, the null hypothesis is equal proportions. If you specify null hypothesis proportions in the TESTP= option, the LRCHISQ option produces a design-based likelihood ratio test for the specified proportions.

By default for one-way tables, and for first-order tests for two-way tables, the design correction is computed from proportion estimates. If you specify LRCHISQ(MODIFIED), the design correction is computed from null hypothesis proportions. For second-order tests for two-way tables, the design correction is always computed from null hypothesis proportions.

You can specify the following *options*:

FIRSTORDER

requests a first-order Rao-Scott likelihood ratio test. This is the default for the LRCHISQ option; if you do not specify LRCHISQ(SECONDORDER), the procedure provides a first-order Rao-Scott test.

MODIFIED

uses the null hypothesis proportions to compute the Rao-Scott design correction. By default (if you do not specify LRCHISQ(MODIFIED)), the procedure uses proportion estimates to compute the design correction for all first-order tests and for second-order tests for one-way tables. For second-order tests for two-way tables, the procedure always uses null hypothesis proportions to compute the design correction.

SECONDORDER

requests a second-order (Satterthwaite) Rao-Scott likelihood ratio test. For more information, see the section “Rao-Scott Likelihood Ratio Chi-Square Test” on page 8033.

NOCELLPERCENT

suppresses the display of overall cell percentages in the crosstabulation table, as well as the standard errors of the percentages. The NOCELLPERCENT option does not suppress the display of row or column percentages, which you can request by specifying the ROW or COLUMN option.

NOFREQ

suppresses the display of cell frequencies in the crosstabulation table. The NOFREQ option also suppresses the display of row, column, and overall table frequencies.

NOPERCENT

suppresses the display of all percentages in the crosstabulation table. The NOPERCENT option also suppresses the display of standard errors of the percentages. Use the NOCELLPERCENT option to suppress display of overall cell percentages but allow display of row or column percentages.

NOPRINT

suppresses the display of frequency and crosstabulation tables but displays all requested statistical tests. This option disables the Output Delivery System (ODS) for the suppressed tables. For more information, see Chapter 20, "Using the Output Delivery System."

NOSPARSE

suppresses the display of variable levels with zero frequency in two-way tables. By default, the procedure displays all levels of the column variable within each level of the row variable, including any column variable levels with zero frequency for that row. For multiway tables, the procedure displays all levels of the row variable for each layer of the table by default, including any row variable levels with zero frequency for the layer.

NOSTD

suppresses the display of all standard errors in the crosstabulation table.

NOTOTAL

suppresses the display of row totals, column totals, and overall totals in the crosstabulation table.

NOWT

suppresses the display of weighted frequencies in the crosstabulation table. The NOWT option also suppresses the display of standard errors of the weighted frequencies.

OR

RELRISK

requests estimates of the odds ratio, the column 1 relative risk, and the column 2 relative risk for 2×2 tables. The OR option also provides confidence limits for these statistics. For more information, see the section "Odds Ratio and Relative Risks" on page 8023.

You can specify the confidence level in the ALPHA= option. By default, ALPHA=0.05, which produces 95% confidence limits.

PEARSONRES

displays each crosstabulation table cell's Pearson residual, which is the square root of the table cell's contribution to the Pearson chi-square statistic. The Pearson residual is computed as $(\textit{weighted frequency} - \textit{expected}) \ / \ \sqrt{\textit{expected}}$, where *weighted frequency* is the weighted frequency of the table cell and *expected* is the expected weighted frequency, which is computed under the null hypothesis that the row and column variables are independent. You can display the expected values, the deviations, and the cell chi-squares by specifying the EXPECTED, DEVIATION, and CELLCHI2 options, respectively. For more information, see the sections "Expected Weighted Frequency" on page 8021 and "Rao-Scott Chi-Square Test" on page 8028. This option has no effect for one-way tables.

PLOTS < (*global-plot-options*) > < =*plot-request* < (*plot-options*) > >

PLOTS < (*global-plot-options*) >

< =(*plot-request* < (*plot-options*) > < ...*plot-request* < (*plot-options*) > >) >

controls the plots that are produced through ODS Graphics. *Plot-requests* identify the plots, and *plot-options* control the appearance and content of the plots. You can specify *plot-options* in parentheses after a *plot-request*. A *global-plot-option* applies to all plots for which it is available unless it is altered by a specific *plot-option*. You can specify *global-plot-options* in parentheses after the PLOTS option.

When you specify only one *plot-request*, you can omit the parentheses around the *plot-request*. For example:

```
plots=all
plots=wtfreqplot
plots=(wtfreqplot oddsratioplot)
plots(only)=(riskdiffplot relriskplot)
```

ODS Graphics must be enabled before plots can be requested. For example:

```
ods graphics on;
proc surveyfreq;
   tables treatment*response / chisq plots=wtfreqplot;
   weight wt;
run;
ods graphics off;
```

For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 606 in Chapter 21, "Statistical Graphics Using ODS."

If ODS Graphics is enabled but you do not specify the PLOTS= option, PROC SURVEYFREQ produces all plots that are associated with the analyses that you request, with the exception of weighted frequency plots and mosaic plots. To produce a weighted frequency plot or mosaic plot when ODS Graphics is enabled, you must specify the WTFREQPLOT or MOSAICPLOT *plot-request*, or you must specify the PLOTS=ALL option. PROC SURVEYFREQ produces the remaining plots (listed in Table 97.5) by default when you request the corresponding TABLES statement options.

You can suppress default plots and request specific plots by using the PLOTS(ONLY)= option; PLOTS(ONLY)=(*plot-requests*) produces only the plots that are specified as *plot-requests*. You can suppress all plots by specifying the PLOTS=NONE option.

See Figure 97.4 and Figure 97.7 for examples of plots that PROC SURVEYFREQ produces. For information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS."

Plot Requests

Table 97.5 lists the available *plot-requests* together with their required TABLES statement options. Descriptions of the *plot-requests* follow the table in alphabetical order.

Table 97.5 Plot Requests

Plot Request	Description	Required TABLES Statement Option
ALL	All plots	None
KAPPAPLOT	Kappa plot	AGREE or KAPPA ($h \times r \times r$ table)
MOSAICPLOT	Mosaic plot	Crosstabulation table request
NONE	No plots	None
ODDSRATIOPLOT	Odds ratio plot	OR ($h \times 2 \times 2$ table)
RELRISKPLOT	Relative risk plot	OR ($h \times 2 \times 2$ table)
RISKDIFFPLOT	Risk difference plot	RISK, RISK1, or RISK2 ($h \times 2 \times 2$ table)
WTFREQPLOT	Weighted frequency plot	Frequency or crosstabulation table request
WTKAPPAPLOT	Weighted kappa plot	AGREE or WTKAPPA ($h \times r \times r$ table, $r > 2$)

The following *plot-requests* are available:

ALL

requests all plots that are associated with the specified analyses. If you specify the PLOTS=ALL option, PROC SURVEYFREQ also produces the weighted frequency and mosaic plots that are associated with the tables that you request. (PROC SURVEYFREQ does not produce weighted frequency and mosaic plots by default when ODS Graphics is enabled.)

KAPPAPLOT < (*plot-options*) >

requests a plot of kappa coefficients with confidence limits. Kappa plots are available for multiway square tables and display the simple kappa coefficient (with confidence limits) for each two-way table layer. To produce a kappa plot, you must specify the KAPPA or AGREE option in the TABLES statement to compute kappa coefficients.

Table 97.6 lists the *plot-options* that are available for kappa plots. For descriptions of the *plot-options*, see the subsection "Plot Options."

Table 97.6 Plot Options for KAPPAPLOT and WTKAPPAPLOT

Plot Option	Description	Values
CLDISPLAY=	Error bar type	**BAR**, **LINE**, **LINEARROW**, **SERIF***, or **SERIFARROW**
NPANELPOS=	Statistics per graphic	Number (All*)
ORDER=	Order of two-way levels	**ASCENDING** or **DESCENDING**
RANGE=	Range to display	Values or **CLIP**
STATS	Statistic values	None

*Default

MOSAICPLOT < (*plot-options*) >

requests a mosaic plot. Mosaic plots are available for crosstabulation tables. For multiway tables, PROC SURVEYFREQ provides a mosaic plot for each two-way table layer.

To produce a mosaic plot, you must specify the MOSAICPLOT *plot-request* in the PLOTS= option, or you must specify the PLOTS=ALL option. PROC SURVEYFREQ does not produce mosaic plots by default when ODS Graphics is enabled.

Mosaic plots display tiles that correspond to the crosstabulation table cells. The areas of the tiles are proportional to the weighted frequencies of the table cells. The column variable is displayed on the X axis, and the tile widths are proportional to the relative weighted frequencies of the column variable levels. The row variable is displayed on the Y axis, and the tile heights are proportional to the relative weighted frequencies of the row levels within column levels. For more information, see Friendly (2000).

By default, the colors of the tiles correspond to the row variable levels. If you specify the COLORSTAT *plot-option*, the tiles are colored according to the values of the Pearson residuals.

You can specify the following *plot-options*:

COLORSTAT < =PEARSONRES >

colors the mosaic plot tiles according to the values of the Pearson residuals. A table cell's Pearson residual is the square root of its contribution to the Pearson chi-square statistic. The Pearson residual is computed as $(\textit{weighted frequency} - \textit{expected}) \; / \; \sqrt{\textit{expected}}$, where *weighted frequency* is the weighted frequency of the table cell and *expected* is the expected weighted frequency. You can specify the PEARSONRES option to display the Pearson residuals in the crosstabulation table.

SQUARE

produces a square mosaic plot, where the height of the Y axis equals the width of the X axis. In a square mosaic plot, the scale of the relative weighted frequencies is the same on both axes. By default, PROC SURVEYFREQ produces a rectangular mosaic plot.

NONE

suppresses all plots.

ODDSRATIOPLOT < (*plot-options*) >

requests a plot of odds ratios with confidence limits. Odds ratio plots are available for multiway 2×2 tables and display the odds ratio (with confidence limits) for each 2×2 table layer. To produce an odds ratio plot, you must specify the OR option in the TABLES statement for a multiway 2×2 table.

Table 97.7 lists the *plot-options* that are available for odds ratio plots. For descriptions of the *plot-options*, see the subsection "Plot Options."

Table 97.7 Plot Options for ODDSRATIOPLOT, RELRISKPLOT, and RISKDIFFPLOT

Plot Option	Description	Values
CLDISPLAY=	Error bar type	**BAR**, **LINE**, **LINEARROW**, **SERIF***, or **SERIFARROW**
COLUMN=**	Risk column	**1*** or **2**
LOGBASE=***	Axis scale	**2**, **E**, or **10**
NPANELPOS=	Statistics per graphic	Number (All*)
ORDER=	Order of two-way levels	**ASCENDING** or **DESCENDING**
RANGE=	Range to display	Values or **CLIP**
STATS	Statistic values	None

*Default

**Available for RELRISKPLOT and RISKDIFFPLOT

***Available for ODDSRATIOPLOT and RELRISKPLOT

RELRISKPLOT < (*plot-options*) >

requests a plot of relative risks with confidence limits. Relative risk plots are available for multiway 2×2 tables and display the relative risk (with confidence limits) for each 2×2 table layer. To produce a relative risk plot, you must specify the OR option in the TABLES statement for a multiway 2×2 table.

Table 97.7 lists the *plot-options* that are available for relative risk plots. For descriptions of the *plot-options*, see the subsection "Plot Options."

RISKDIFFPLOT < (*plot-options*) >

requests a plot of risk differences with confidence limits. Risk difference plots are available for multiway 2×2 tables and display the risk difference (with confidence limits) for each 2×2 table layer. To produce a risk difference plot, you must specify the RISK, RISK1, or RISK2 option in the TABLES statement for a multiway 2×2 table.

Table 97.7 lists the *plot-options* that are available for risk difference plots. For descriptions of the *plot-options*, see the subsection "Plot Options."

WTFREQPLOT < (*plot-options*) >

requests a weighted frequency plot. Weighted frequency plots are available for frequency and crosstabulation tables. For multiway tables, PROC SURVEYFREQ provides a two-way weighted frequency plot for each two-way table layer.

To produce a weighted frequency plot, you must specify the WTFREQPLOT *plot-request* in the PLOTS= option, or you must specify the PLOTS=ALL option. PROC SURVEYFREQ does not produce weighted frequency plots by default when ODS Graphics is enabled.

By default, PROC SURVEYFREQ displays weighted frequency plots as bar charts. You can specify the TYPE=DOTPLOT *plot-option* to display frequency plots as dot plots. You can plot weighted percentages instead of frequencies by specifying the SCALE=PERCENT *plot-option*. There are four frequency plot layouts available, which you can request by specifying the TWOWAY= *plot-option*. For more information, see the subsection "Plot Options."

By default, the primary grouping of graph cells in a two-way layout is by column variable. Row variable levels are then displayed within column variable levels. You can specify the GROUPBY=ROW *plot-option* to group first by row variable.

Weighted frequency plots for one-way tables display confidence limits by default. For two-way tables, weighted frequency plots display confidence limits by default in the TWOWAY=GROUPVERTICAL and TWOWAY=GROUPHORIZONTAL layouts. You can suppress confidence limits by specifying the CLBAR=NO *plot-option*. Confidence limits are not available for two-way plots in the TWOWAY=CLUSTER and TWOWAY=STACKED layouts.

Table 97.8 lists the *plot-options* that are available for weighted frequency plots. For descriptions of the *plot-options*, see the subsection "Plot Options."

Table 97.8 Plot Options for WTFREQPLOT

Plot Option	Description	Values
CLBAR=	Confidence limit bars	**NO** or **YES***
GROUPBY=**	Primary group	**COLUMN*** or **ROW**
NPANELPOS=**	Sections per panel	Number (4*)
ORIENT=	Orientation	**HORIZONTAL** or **VERTICAL***
SCALE=	Scale	**PERCENT** or **WTFREQ***
TWOWAY=**	Two-way layout	**CLUSTER**, **GROUPHORIZONTAL**, **GROUPVERTICAL***, or **STACKED**
TYPE=	Type	**BARCHART*** or **DOTPLOT**

*Default

**For two-way tables

WTKAPPAPLOT < (*plot-options***) >**

requests a plot of weighted kappa coefficients with confidence limits. Weighted kappa plots are available for multiway square tables and display the weighted kappa coefficient (with confidence limits) for each two-way table layer. To produce a weighted kappa plot, you must specify the WTKAPPA or AGREE option in the TABLES statement to compute weighted kappa coefficients, and the table dimension must be greater than 2.

Table 97.6 lists the *plot-options* that are available for weighted kappa plots. For descriptions of the *plot-options*, see the subsection "Plot Options."

Global Plot Options

A *global-plot-option* applies to all plots for which the option is available unless it is altered by an individual *plot-option*. All *plot-options* that are listed in Table 97.8 and Table 97.7 are available as *global-plot-options*. The ONLY option is also available as a *global-plot-option*.

You can specify *global-plot-options* in parentheses after the PLOTS option. For example:

```
plots(order=ascending stats)=(riskdiffplot oddsratioplot)
plots(only)=wtfreqplot
```

In addition to the *plot-options* that are listed in Table 97.8 and Table 97.7, you can specify the following *global-plot-option* in parentheses after the PLOTS option:

ONLY

suppresses the default plots and requests only the plots that are specified as *plot-requests*.

Plot Options

You can specify the following *plot-options* in parentheses after a *plot-request*.

CLBAR=NO | YES

controls the confidence limit error bars in weighted frequency plots (WTFREQPLOT). By default, CLBAR=YES, which displays confidence limits error bars; CLBAR=NO suppresses confidence limit error bars.

This *plot-option* applies to all weighted frequency plots except those two-way plots that are displayed in the TWOWAY=CLUSTER or TWOWAY=STACKED layout. Confidence limit error bars are not available in the TWOWAY=CLUSTER and TWOWAY=STACKED layouts.

CLDISPLAY=BAR < *width* > | LINE | LINEARROW | SERIF | SERIFARROW

controls the appearance of the confidence limit error bars. This *plot-option* is available for the following plots: KAPPAPLOT, ODDSRATIOPLOT, RELRISKPLOT, RISKDIFFPLOT, and WTKAPPAPLOT.

The default is CLDISPLAY=SERIF, which displays the confidence limits as lines with serifs. CLDISPLAY=LINE displays the confidence limits as plain lines without serifs. The CLDISPLAY=SERIFARROW and CLDISPLAY=LINEARROW *plot-options* display arrowheads on any error bars that are clipped by the RANGE= *plot-option*; if an entire error bar is cut from the plot, the plot displays an arrowhead that points toward the statistic.

CLDISPLAY=BAR displays the confidence limits as bars. By default, the width of the bars equals the size of the marker for the estimate. You can control the width of the bars and the size of the marker by specifying the value of *width* as a percentage of the distance between bars, $0 < \mathit{width} \leq 1$. The bar might disappear when the value of *width* is very small.

COLUMN=1 | 2

specifies the 2×2 table column to use to compute the risk (proportion). This *plot-option* is available for the relative risk plot (RELRISKPLOT) and the risk difference plot (RISKDIFFPLOT). If you specify COLUMN=1, the plot displays the column 1 relative risks or the column 1 risk differences. Similarly, if you specify COLUMN=2, the plot displays the column 2 relative risks or risk differences.

For relative risk plots, the default is COLUMN=1. For risk difference plots, the default if COLUMN=1 if you request computation of both column 1 and column 2 risk differences with the RISK option. If you request computation of only column 1 (or only column 2) risk differences by specifying the RISK1 (or RISK2) option, by default the risk difference plot displays these risk differences.

GROUPBY=COLUMN | ROW

specifies the primary grouping for two-way weighted frequency plots, which you can request by specifying the WTFREQPLOT *plot-request*.

The default is GROUPBY=COLUMN, which groups graph cells first by column variable and displays row variable levels within column variable levels. You can specify GROUPBY=ROW to group first by row variable. In two-way and multiway table requests, the column variable is the last variable specified and forms the columns of the crosstabulation table. The row variable is the next-to-last variable specified and forms the rows of the table.

By default for a bar chart that is displayed in the TWOWAY=STACKED layout, bars correspond to the column variable levels and row levels are displayed (stacked) within each column bar. By default for a bar chart that is displayed in the TWOWAY=CLUSTER layout, bars are first grouped by column variable levels, and row levels are displayed as adjacent bars within each column-level group. You can reverse the default row and column variable groupings by specifying GROUPBY=ROW.

LOGBASE=2 | E | 10

applies to the odds ratio plot (ODDSRATIOPLOT) and the relative risk plot (RELRISKPLOT). This *plot-option* displays the odds ratio or relative risk axis on the log scale that you specify.

NPANELPOS=*n*

divides the plot into multiple panels that display at most $|n|$ statistics or sections.

If n is positive, the number of statistics or sections per panel is balanced; if n is negative, the number of statistics per panel is not balanced. For example, suppose you want to display 21 odds ratios. NPANELPOS=20 displays two panels, the first with 11 odds ratios and the second with 10 odds ratios; NPANELPOS=–20 displays 20 odds ratios in the first panel but only 1 in the second panel. This *plot-option* is available for all plots except mosaic plots and one-way weighted frequency plots.

For two-way weighted frequency plots (WTFREQPLOT), NPANELPOS=n requests that panels display at most $|n|$ sections, where sections correspond to row or column variable levels, depending on the type of plot and the grouping. By default, n=4 and each panel includes at most four sections. This *plot-option* applies to two-way plots that are displayed in the TWOWAY=GROUPVERTICAL or TWOWAY=GROUPHORIZONTAL layout. The NPANELPOS= *plot-option* does not apply to the TWOWAY=CLUSTER and TWOWAY=STACKED layouts, which are always displayed in a single panel.

For plots that display statistics with confidence limits, NPANELPOS=n requests that panels display at most $|n|$ statistics. By default, n=0 and all statistics are displayed in a single panel. This *plot-option* applies to the following plots: KAPPAPLOT, ODDSRATIOPLOT, RELRISKPLOT, RISKDIFFPLOT, and WTKAPPAPLOT.

ORDER=ASCENDING | DESCENDING

displays the two-way table (layer) statistics in order of the statistic value. If you specify ORDER=ASCENDING or ORDER=DESCENDING, the plot displays the statistics in ascending or descending order, respectively. By default, the order of the statistics in the plot matches the order that the two-way table layers appear in the multiway table.

This *plot-option* is available for the following plots: KAPPAPLOT, ODDSRATIOPLOT, RELRISKPLOT, RISKDIFFPLOT, and WTKAPPAPLOT.

ORIENT=HORIZONTAL | VERTICAL

controls the orientation of weighted frequency plots (WTFREQPLOT). This *plot-option* places the variable levels on the Y axis and the weighted frequencies or percentages on the X axis. ORIENT=VERTICAL places the variable levels on the X axis. The default orientation is ORIENT=VERTICAL for bar charts (TYPE=BARCHART) and ORIENT=HORIZONTAL for dot plots (TYPE=DOTPLOT).

RANGE=(< *min* >< , *max* >)| CLIP

specifies the range of values to display. If you specify RANGE=CLIP, the confidence limits are clipped and the display range is determined by the minimum and maximum values of the estimates. By default, the display range includes all confidence limits.

This *plot-option* is available for the following plots: KAPPAPLOT, ODDSRATIOPLOT, RELRISKPLOT, RISKDIFFPLOT, and WTKAPPAPLOT.

SCALE=PERCENT | WTFREQ

specifies the scale of the frequencies in weighted frequency plots (WTFREQPLOT). SCALE=WTFREQ displays weighted frequencies (totals), and SCALE=PERCENT displays percentages. The default scale is SCALE=WTFREQ.

STATS

displays the values of the statistics and their confidence limits on the right side of the plot. If you do not specify this *plot-option*, the statistic values are not displayed.

This *plot-option* is available for the following plots: KAPPAPLOT, ODDSRATIOPLOT, RELRISKPLOT, RISKDIFFPLOT, and WTKAPPAPLOT.

TWOWAY=CLUSTER | GROUPHORIZONTAL | GROUPVERTICAL | STACKED

specifies the layout for two-way weighted frequency plots (WTFREQPLOT).

All TWOWAY= layouts are available for bar charts (TYPE=BARCHART). All TWOWAY= layouts except TWOWAY=CLUSTER are available for dot plots (TYPE=DOTPLOT). Confidence limits (CLBAR=) can be displayed in the GROUPVERTICAL and GROUPHORIZONTAL layouts. Confidence limits are not available in the STACKED and CLUSTER layouts. The ORIENT= and GROUPBY= *plot-options* are available for all TWOWAY= layouts.

The default two-way layout is TWOWAY=GROUPVERTICAL, which produces a grouped plot that has a vertical common baseline. By default for bar charts (TYPE=BARCHART, ORIENT=VERTICAL), the X axis displays column variable levels, and the Y axis displays weighted frequencies. The plot includes a vertical (Y-axis) block for each row variable level. The relative positions of the graph cells in this plot layout are the same as the relative positions of the table cells in the crosstabulation table. You can reverse the default row and column grouping by specifying the GROUPBY=ROW *plot-option*.

The TWOWAY=GROUPHORIZONTAL layout produces a grouped plot that has a horizontal common baseline. By default (GROUPBY=COLUMN), the plot displays a block on the X axis for each column variable level. Within each column-level block, the plot displays row variable levels.

The TWOWAY=STACKED layout produces stacked displays of weighted frequencies. By default (GROUPBY=COLUMN) in a stacked bar chart, the bars correspond to column variable levels, and row levels are stacked within each column level. By default in a stacked dot plot, the dotted lines correspond to column levels, and cell weighted frequencies are plotted as data dots on the corresponding column line. The dot color identifies the row level.

The TWOWAY=CLUSTER layout, which is available only for bar charts, displays groups of adjacent bars. By default, the primary grouping is by column variable level, and row levels are displayed within each column level.

You can reverse the default row and column grouping in any layout by specifying the GROUPBY=ROW *plot-option*. The default is GROUPBY=COLUMN, which groups first by column variable.

TYPE=BARCHART | DOTPLOT

specifies the type (form) of the weighted frequency plots (WTFREQPLOT). TYPE=BARCHART produces a bar chart and TYPE=DOTPLOT produces a dot plot. The default type is TYPE=BARCHART.

RISK

RISKDIFF

requests risk statistics for 2×2 tables. The RISK option also provides standard errors and confidence limits for these statistics. Risk statistics include the row 1 risk (proportion), row 2 risk, overall risk, and risk difference. For more information, see the section "Risks and Risk Difference" on page 8022.

The RISK option provides both column 1 and column 2 risks. To request only column 1 or column 2 risks, use the RISK1 or RISK2 option.

You can specify the confidence level in the ALPHA= option. By default, ALPHA=0.05, which produces 95% confidence limits.

RISK1

RISKDIFF1

requests column 1 risk statistics for 2×2 tables, together with their standard errors and confidence limits. Risk statistics include the row 1 risk (proportion), row 2 risk, overall risk, and risk difference. For more information, see the section "Risks and Risk Difference" on page 8022.

You can specify the confidence level in the ALPHA= option. By default, ALPHA=0.05, which produces 95% confidence limits.

RISK2

RISKDIFF2

requests column 2 risk statistics for 2×2 tables, together with their standard errors and confidence limits. Risk statistics include the row 1 risk (proportion), row 2 risk, overall risk, and risk difference. For more information, see the section "Risks and Risk Difference" on page 8022.

You can specify the confidence level in the ALPHA= option. By default, ALPHA=0.05, which produces 95% confidence limits.

ROW < (*option*) >

displays the row percentage (estimated proportion of the row total) for each cell in a two-way table. The ROW option also provides the standard errors of the row percentages. For more information, see the section "Row and Column Proportions" on page 8010. This option has no effect for one-way tables.

You can specify the following *option*:

DEFF

displays the design effect for each row percentage in the crosstabulation table. For more information, see the section "Design Effect" on page 8020.

TESTP=(*values*)

specifies null hypothesis proportions (test percentages) for chi-square tests for one-way tables (goodness-of-fit tests). You can separate *values* with blanks or commas, and you must enclose the list of values in parentheses. Specify *values* in probability form as numbers between 0 and 1, where the proportions sum to 1. Or specify *values* in percentage form as numbers between 0 and 100, where the percentages sum to 100. PROC SURVEYFREQ treats the value 1 as the percentage form 1%. The number of TESTP= values must equal the number of variable levels in the one-way table. List these values in the same order in which the corresponding variable levels appear in the output.

When you specify the TESTP= option, PROC SURVEYFREQ displays the specified test percentages in the one-way frequency table. The TESTP= option has no effect for two-way tables.

PROC SURVEYFREQ uses the TESTP= values for the one-way Rao-Scott chi-square test (CHISQ) and for the one-way Rao-Scott likelihood ratio chi-square test (LRCHISQ). See the sections "Rao-Scott Chi-Square Test" on page 8028 and For more information, see the section "Rao-Scott Likelihood Ratio Chi-Square Test" on page 8033.

VAR

displays the variance estimate for each percentage in the crosstabulation table. For more information, see the section "Proportions" on page 8008. By default, PROC SURVEYFREQ displays the standard errors of the percentages.

VARWT

displays the variance estimate for each weighted frequency, or estimated total, in the crosstabulation table. For more information, see the section "Totals" on page 8006. By default, PROC SURVEYFREQ displays the standard deviations of the weighted frequencies.

WCHISQ

requests the Wald chi-square test for two-way tables. For more information, see the section "Wald Chi-Square Test" on page 8035.

WLLCHISQ

requests the Wald log-linear chi-square test for two-way tables. For more information, see the section "Wald Log-Linear Chi-Square Test" on page 8036.

WTFREQ

displays totals (weighted frequencies) and their standard errors when you do not specify a WEIGHT or REPWEIGHTS statement. By default, PROC SURVEYFREQ displays the weighted frequencies only when you specify a WEIGHT or REPWEIGHTS statement. When you do not specify a WEIGHT or REPWEIGHTS statement, PROC SURVEYFREQ assigns all observations a weight of one.

WTKAPPA < (*options*) >

requests the weighted kappa coefficient with its standard error and confidence limits. Weighted kappa coefficients can be computed for square two-way tables, where the number of rows equals the number of columns. For 2×2 tables, the weighted kappa coefficient equals the simple kappa coefficient, and PROC SURVEYFREQ displays only the simple kappa coefficient. For more information, see the section "Weighted Kappa Coefficient" on page 8026.

Weighted kappa coefficients are available when you specify variance estimation by the jackknife method (VARMETHOD=JACKKNIFE) or by balanced repeated replication (VARMETHOD=BRR); weighted kappa coefficients are not available with the Taylor series method (VARMETHOD=TAYLOR).

The weighted kappa coefficient is computed by using agreement weights that reflect the relative agreement between pairs of variable levels. Agreement weights are not the same as *sampling weights*, which you provide by specifying the WEIGHT statement. PROC SURVEYFREQ uses the sampling weights to compute both the simple kappa and weighted kappa coefficients. For more information, see the section “Weighted Kappa Coefficient” on page 8026.

You can specify the confidence level in the ALPHA= option. By default, ALPHA=0.05, which produces 95% confidence limits.

You can specify the following *options*:

PRINTKWTS

displays the agreement weights that PROC SURVEYFREQ uses to compute the weighted kappa coefficient. Agreement weights reflect the relative agreement between pairs of variable levels. By default, PROC SURVEYFREQ computes these weights by using the Cicchetti-Allison form. If you specify the WT=FC option, the procedure uses the Fleiss-Cohen form of agreement weights. For more information, see the section “Weighted Kappa Coefficient” on page 8026.

WT=FC

requests Fleiss-Cohen agreement weights for the weighted kappa computation. By default, PROC SURVEYFREQ uses Cicchetti-Allison agreement weights to compute the weighted kappa coefficient. Agreement weights reflect the relative agreement between pairs of variable levels. For more information, see the section “Weighted Kappa Coefficient” on page 8026.

WEIGHT Statement

WEIGHT *variable* **;**

The WEIGHT statement names the *variable* that contains the sampling weights. This variable must be numeric, and the sampling weights must be positive numbers. If an observation has a weight that is nonpositive or missing, then the procedure omits that observation from the analysis. For more information, see the section “Missing Values” on page 8001. If you specify more than one WEIGHT statement, the procedure uses only the first WEIGHT statement and ignores the rest.

If you do not specify a WEIGHT statement but provide replicate weights by specifying a REPWEIGHTS statement, PROC SURVEYFREQ uses the average of each observation’s replicate weights as the observation’s weight.

If you do not specify a WEIGHT statement or a REPWEIGHTS statement, PROC SURVEYFREQ assigns all observations a weight of one.

Details: SURVEYFREQ Procedure

Specifying the Sample Design

PROC SURVEYFREQ produces tables and statistics that are based on the sample design used to obtain the survey data. PROC SURVEYFREQ can be used for single-stage or multistage designs, with or without stratification, and with or without unequal weighting. To analyze your survey data with PROC SURVEYFREQ, you need to provide sample design information for the procedure. This information can include design strata, clusters, and sampling weights. You can provide sample design information by specifying the STRATA, CLUSTER, and WEIGHT statements and the RATE= or TOTAL= option in the PROC SURVEYFREQ statement.

If you provide replicate weights for BRR or jackknife variance estimation, you do not need to specify a STRATA or CLUSTER statement. Otherwise, you should specify STRATA and CLUSTER statements whenever your design includes stratification and clustering.

When there are clusters (PSUs) in the sample design, the procedure estimates the variance by using the PSUs, as described in the section "Statistical Computations" on page 8004. For a multistage sample design, the variance estimation depends only on the first stage of the sample design. Therefore, the required input includes only first-stage cluster (PSU) and first-stage stratum identification. You do not need to input design information about any additional stages of sampling.

Stratification

If your sample design is stratified at the first stage of sampling, use the STRATA statement to name the variables that form the strata. The combinations of categories of STRATA variables define the strata in the sample, where strata are nonoverlapping subgroups that were sampled independently. If your sample design has stratification at multiple stages, you should identify only the first-stage strata in the STRATA statement.

If you use a REPWEIGHTS statement to provide replicate weights for BRR or jackknife variance estimation, you do not need to specify a STRATA statement. Otherwise, you should specify a STRATA statement whenever your design includes stratification. If you do not specify a STRATA statement or a REPWEIGHTS statement, then PROC SURVEYFREQ assumes there is no stratification at the first stage.

Clustering

If your sample design selects clusters at the first stage of sampling, use the CLUSTER statement to name the variables that identify the first-stage clusters, which are also called primary sampling units (PSUs). The combinations of categories of CLUSTER variables define the clusters in the sample. If there is a STRATA statement, clusters are nested within strata. If your sample design has clustering at multiple stages, you should specify only the first-stage clusters (PSUs) in the CLUSTER statement. PROC SURVEYFREQ assumes that each cluster that is defined by the CLUSTER statement variables represents a PSU in the sample, and that each observation belongs to one PSU.

If you use a REPWEIGHTS statement to provide replicate weights for BRR or jackknife variance estimation, you do not need to specify a CLUSTER statement. Otherwise, you should specify a CLUSTER statement whenever your design includes clustering at the first stage of sampling. If you do not specify a CLUSTER statement, then PROC SURVEYFREQ treats each observation as a PSU.

Weighting

If your sample design includes unequal weighting, use the WEIGHT statement to name the variable that contains the sampling weights. Sampling weights must be positive numbers. If an observation has a weight that is nonpositive or missing, the procedure omits that observation from the analysis. For more information, see the section "Missing Values" on page 8001.

If you do not specify a WEIGHT statement but include a REPWEIGHTS statement, PROC SURVEYFREQ uses the average of each observation's replicate weights as the observation's weight. If you do not specify a WEIGHT statement or a REPWEIGHTS statement, PROC SURVEYFREQ assigns all observations a weight of one.

Population Totals and Sampling Rates

To include a finite population correction (*fpc*) in Taylor series variance estimation, you can input either the sampling rate or the population total by using the RATE= or TOTAL= option in the PROC SURVEYFREQ statement. (You cannot specify both of these options in the same PROC SURVEYFREQ statement.) The RATE= and TOTAL= options apply only to Taylor series variance estimation. The procedure does not use a finite population correction for BRR or jackknife variance estimation.

If you do not specify the RATE= or TOTAL= option, Taylor series variance estimation does not include a finite population correction. For fairly small sampling fractions, it is appropriate to ignore this correction. For more information, see Cochran (1977) and Kish (1965).

If your design has multiple stages of selection and you are specifying the RATE= option, you should input the first-stage sampling rate, which is the ratio of the number of PSUs in the sample to the total number of PSUs in the study population. If you are specifying the TOTAL= option for a multistage design, you should input the total number of PSUs in the study population.

For a nonstratified sample design, or for a stratified sample design with the same sampling rate or the same population total in all strata, you can use the RATE=*value* or TOTAL=*value* option. If your sample design is stratified with different sampling rates or population totals in different strata, use the RATE=*SAS-data-set* or TOTAL=*SAS-data-set* option to name a SAS data set that contains the stratum sampling rates or totals. This data set is called a *secondary data set*, as opposed to the *primary data set* that you specify in the DATA= option.

The secondary data set must contain all the stratification variables listed in the STRATA statement and all the variables in the BY statement. Furthermore, the BY groups must appear in the same order as in the primary data set. If there are formats that are associated with the STRATA variables and the BY variables, then the formats must be consistent in the primary and the secondary data sets. If you specify the TOTAL=*SAS-data-set* option, the secondary data set must have a variable named _TOTAL_ that contains the stratum population totals. If you specify the RATE=*SAS-data-set* option, the secondary data set must have a variable named _RATE_ that contains the stratum sampling rates. If the secondary data set contains more than one observation for any one stratum, the procedure uses the first value of _TOTAL_ or _RATE_ for that stratum and ignores the rest.

The *value* in the RATE= option or the values of _RATE_ in the secondary data set must be nonnegative numbers. You can specify *value* as a number between 0 and 1. Or you can specify *value* in percentage form as a number between 1 and 100, and PROC SURVEYFREQ converts that number to a proportion. The procedure treats the value 1 as 100% instead of 1%.

If you specify the TOTAL=*value* option, *value* must not be less than the sample size. If you provide stratum population totals in a secondary data set, these values must not be less than the corresponding stratum sample sizes.

Domain Analysis

PROC SURVEYFREQ provides domain analysis through its multiway table capability. *Domain analysis* refers to the computation of statistics for domains (subpopulations), in addition to the computation of statistics for the entire study population. Formation of subpopulations can be unrelated to the sample design, and so the domain sample sizes can actually be random variables. Domain analysis takes this variability into account by using the entire sample to estimate the variance of domain estimates. Domain analysis is also known as subgroup analysis, subpopulation analysis, and subdomain analysis. For more information about domain analysis, see Lohr (2010), Cochran (1977), Fuller et al. (1989).

To request domain analysis, you should include the domain variable(s) in your TABLES statement request. For example, specifying DOMAIN * A * B in a TABLES statement produces separate two-way tables of A by B for each level of DOMAIN. If your domains are formed by more than one variable, you can specify DomainVariable_1 * DomainVariable_2 * A * B, for example, to obtain two-way tables of A by B for each domain formed by the different combinations of levels for DomainVariable_1 and DomainVariable_2. See Example 97.2 for an example of domain analysis.

If you specify DOMAIN * A in a TABLES statement, the values of the variable DOMAIN form the table rows. The two-way table lists levels of the variable A within each level of the row variable DOMAIN. Specify the ROW option in the TABLES statement to obtain the row percentages and their standard errors. This provides the one-way distribution of A for each domain (level of the variable DOMAIN).

Including the domain variables in a TABLES statement request provides a different analysis from the analysis that you obtain by using a BY statement; a BY statement provides completely separate analyses of the BY groups. You can use the BY statement to analyze the data set by subgroups, but it is critical to note that this does *not* produce a valid domain analysis. The BY statement is appropriate only when the number of units in each subgroup is known with certainty. For example, you can use a BY statement to obtain stratum level estimates when the stratum sample sizes are fixed. When the subgroup sample size is a random variable, you should include the domain variables in your TABLES statement request.

Missing Values

WEIGHT Variable

If an observation has a missing or nonpositive value for the WEIGHT variable, PROC SURVEYFREQ excludes that observation from the analysis.

REPWEIGHTS Variables

If you provide replicate weights by specifying a REPWEIGHTS statement, all REPWEIGHTS variable values must be nonmissing. Similarly, if you provide jackknife coefficients by specifying the JKCOEFS= option in the REPWEIGHTS statement, all values of the JKCoefficient variable must be nonmissing. If any replicate weight or jackknife coefficient is missing, PROC SURVEYFREQ does not perform the analysis.

STRATA and CLUSTER Variables

If an observation has a missing value for any STRATA or CLUSTER variable, PROC SURVEYFREQ excludes that observation from the analysis unless you specify the MISSING option in the PROC

SURVEYFREQ statement. If you specify the MISSING option, PROC SURVEYFREQ treats missing values as a valid (nonmissing) category for all categorical variables, which include STRATA, CLUSTER, and TABLES variables.

TABLES Variables

If an observation has a missing value for any variable in the TABLES request, PROC SURVEYFREQ excludes that observation from the crosstabulation table (and all associated analyses) unless you specify the MISSING or NOMCAR option in the PROC SURVEYFREQ statement. When the procedure excludes observations with missing values from a table, it displays the total frequency of missing observations below the table.

If you specify the MISSING option, PROC SURVEYFREQ treats missing values as a valid (nonmissing) level for each TABLES variable. The procedure displays these levels in the crosstabulation table and includes them in the computation of totals, percentages, and all other table statistics.

If you specify the NOMCAR option in the PROC SURVEYFREQ statement for Taylor series variance estimation, the procedure includes observations with missing values of TABLES variables in the variance computations. The NOMCAR option does not display missing levels in the crosstabulation table or compute percentages and totals for missing levels.

The NOMCAR Option

The NOMCAR option in the PROC SURVEYFREQ statement includes observations with missing values of TABLES variables in the variance computations as *not missing completely at random* (NOMCAR) for Taylor series variance estimation. By default, observations are completely excluded from the analysis if they have missing values for any of the variables in the current TABLES request. This default treatment is based on the assumption that the values are *missing completely at random* (MCAR), and assumes that the analysis results should not be substantially different between the missing and nonmissing groups. For more information, see the section “Analysis Considerations” on page 8003.

When you specify the NOMCAR option, PROC SURVEYFREQ computes variance estimates by analyzing the nonmissing values as a domain (subpopulation), where the entire population includes both nonmissing and missing domains.

The NOMCAR option has no effect when you specify the MISSING option, which treats missing values as a valid nonmissing level. The NOMCAR option does not affect the inclusion of observations with missing values of the WEIGHT, CLUSTER, or STRATA variables. Observations with missing values of the WEIGHT variable are always excluded from the analysis. Observations with missing values of the CLUSTER or STRATA variables are excluded unless you specify the MISSING option.

The NOMCAR option applies only to Taylor series variance estimation VARMETHOD=TAYLOR. The replication methods, which you can request by specifying the VARMETHOD=BRR and VARMETHOD=JACKKNIFE options, do not use the NOMCAR option.

Degrees of Freedom

PROC SURVEYFREQ computes the degrees of freedom to obtain the t-percentile for confidence limits for proportions, totals, and other statistics. The procedure also uses the degrees of freedom for the F statistics in the Rao-Scott and Wald chi-square tests. The degrees of freedom computation depends on the sample design and the variance estimation method. For more information, see the section “Degrees of Freedom” on page 8019. Missing values can affect the degrees of freedom computation.

Taylor Series Variance Estimation

The degrees of freedom can depend on the number of clusters, the number of strata, and the number of observations. For Taylor series variance estimation, these numbers are based on the observations included in the analysis of the individual table. These numbers do not count observations that are excluded from the table due to missing values. If all values in a stratum are excluded from the analysis of a table as missing values, then that stratum is called an *empty stratum*. Empty strata are not counted in the total number of strata for the table. Similarly, empty clusters and missing observations are not included in the total counts of clusters and observations that are used to compute the degrees of freedom for the analysis.

If you specify the MISSING option, missing values are treated as valid nonmissing levels and are included in computing degrees of freedom. If you specify the NOMCAR option for Taylor series variance estimation, observations with missing values of the TABLES variables are included in computing degrees of freedom.

Replicate-Based Variance Estimation

For BRR or jackknife variance estimation, by default PROC SURVEYFREQ computes the degrees of freedom by using all valid observations in the input data set. A valid observation is an observation that has a positive value of the WEIGHT variable and nonmissing values of the STRATA and CLUSTER variables unless you specify the MISSING option. For information about valid observations, see the section "Data Summary Table" on page 8038.

If you specify the DFADJ *method-option* for VARMETHOD=BRR or VARMETHOD=JACKKNIFE, the procedure computes the degrees of freedom based on the nonmissing observations included in the individual table request. This excludes any empty strata or clusters that occur when observations with missing values of the TABLES variables are removed from the analysis for that table.

Table Summary Output Data Set

For each table request, PROC SURVEYFREQ produces a nondisplayed ODS table, "Table Summary," which contains the number of (nonmissing) observations, strata, and clusters that are included in the analysis of the individual table. If there are missing observations, empty strata, or empty clusters excluded from the analysis, the "Table Summary" data set also contains this information. If you request any confidence limits or chi-square tests for the table, which require degrees of freedom, the "Table Summary" data set provides the degrees of freedom.

Due to missing values, the number of observations used for an individual table analysis can differ from the number of valid observations in the input data set, which is reported in the "Data Summary" table. Similarly, a difference can also occur for the number of clusters or strata. See Example 97.3 for more information about the "Table Summary" output data set.

If you specify the NOMCAR option for Taylor series variance estimation, the "Table Summary" data set reflects all observations used for variance estimation, which includes those observations with missing values of the TABLES variables.

Analysis Considerations

If you have missing values in your survey data for any reason (such as nonresponse), this can compromise the quality of your survey results. An observation without missing values is called a *complete respondent*, and an observation with missing values is called an *incomplete respondent*. If the complete respondents are different from the incomplete respondents with regard to a survey effect or outcome, then survey estimates will be biased and will not accurately represent the survey population. There are a variety of techniques in sample design and survey operations that can reduce nonresponse. After data collection is complete, you can use imputation to replace missing values with acceptable values, and you can use sampling weight adjustments to compensate for nonresponse. You should complete this data preparation and adjustment before you analyze your data with PROC SURVEYFREQ. For more information, see Cochran (1977), Kalton and Kasprzyk (1986), and Brick and Kalton (1996).

Statistical Computations

Variance Estimation

PROC SURVEYFREQ provides a choice of variance estimation methods for complex survey data. In addition to the Taylor series linearization method, the procedures offer two replication-based (resampling) methods—balanced repeated replication (BRR) and the delete-1 jackknife. These variance estimation methods usually give similar, satisfactory results (Lohr 2010; Särndal, Swensson, and Wretman 1992; Wolter 1985). The choice of a variance estimation method can depend on the sample design used, the sample design information available, the parameters to be estimated, and computational issues. For more information, see Lohr (2010).

Taylor Series Variance Estimation

The Taylor series linearization method can be used to estimate standard errors of proportions and other statistics for crosstabulation tables. For sample survey data, the proportion estimator is a ratio estimator formed from estimators of totals. For example, to estimate the proportion in a crosstabulation table cell, the procedure uses the ratio of the estimator of the cell total frequency to the estimator of the overall population total, where these totals are linear statistics computed from the survey data. The Taylor series expansion method obtains a first-order linear approximation for the ratio estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971; Fuller 1975). For more information about Taylor series variance estimation for sample survey data, see Lohr (2010), Särndal, Swensson, and Wretman (1992), Lee, Forthofer, and Lorimor (1989), and Wolter (1985).

When there are clusters (PSUs) in the sample design, the Taylor series method estimates variance from the variance among PSUs. When the design is stratified, the procedure combines stratum variance estimates to compute the overall variance estimate. For a multistage sample design, the variance estimation depends only on the first stage of the sample design. So the required input includes only first-stage cluster (PSU) and first-stage stratum identification. You do not need to input design information about any additional stages of sampling. This variance estimation method assumes that the first-stage sampling fraction is small, or the first-stage sample is drawn with replacement, as it often is in practice.

For more information about Taylor series variance estimation, see the sections "Proportions" on page 8008, "Row and Column Proportions" on page 8010, "Risks and Risk Difference" on page 8022, and "Odds Ratio and Relative Risks" on page 8023.

Replication-Based Variance Estimation

Replication-based methods for variance estimation draw multiple replicates (subsamples) from the full sample by following a specific resampling scheme. Commonly used resampling schemes include *balanced repeated replication* (BRR) and the *jackknife*. PROC SURVEYFREQ estimates the parameter of interest (a proportion, total, odds ratio, or other statistic) from each replicate, and then uses the variability among replicate estimates to estimate the overall variance of the parameter estimate. For more information, see Wolter (1985) and Lohr (2010).

The BRR variance estimation method requires a stratified sample design with two PSUs in each stratum. Each replicate is obtained by deleting one PSU per stratum according to the corresponding Hadamard matrix and adjusting the original weights of the remaining PSUs. The adjusted weights are called *replicate weights*. PROC SURVEYFREQ also provides Fay's method, which is a modification of the BRR method. For more information, see the section "Balanced Repeated Replication (BRR)" on page 8011.

The jackknife method deletes one PSU at a time from the full sample to create replicates, and modifies the original weights to obtain replicate weights. The total number of replicates equals the number of PSUs. If the sample design is stratified, each stratum must contain at least two PSUs, and the jackknife is applied separately within each stratum. For more information, see the section "The Jackknife Method" on page 8014.

Instead of having PROC SURVEYFREQ generate replicate weights for the analysis, you can input your own replicate weights with a REPWEIGHTS statement. This can be useful if you need to do multiple analyses with the same set of replicate weights, or if you have access to replicate weights instead of design information. For more information, see the section "Replicate Weight Output Data Set" on page 8037.

Definitions and Notation

For a stratified clustered sample design, define the following:

h	$=$	$1, 2, \ldots, H$	is the stratum number, with a total of H strata
i	$=$	$1, 2, \ldots, n_h$	is the cluster number within stratum h, with a total of n_h sample clusters in stratum$h$
j	$=$	$1, 2, \ldots, m_{hi}$	is the unit number within cluster i of stratum h with a total of m_{hi} sample units from cluster i of stratum h
n	$=$	$\sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$	is the total number of observations in the sample

f_h = first-stage sampling rate for stratum h

W_{hij} = sampling weight of unit j in cluster i of stratum h

The sampling rate f_h, which is used in Taylor series variance estimation, is the fraction of first-stage units (PSUs) selected for the sample. You can specify the stratum sampling rates with the RATE= option. Or if you specify population totals with the TOTAL= option, PROC SURVEYFREQ computes f_h as the ratio of stratum sample size to the stratum total, in terms of PSUs. For more information, see the section "Population Totals and Sampling Rates" on page 8000. If you do not specify the RATE= option or the TOTAL= option, then the procedure assumes that the stratum sampling rates f_h are negligible and does not use a finite population correction when computing variances.

This notation is also applicable to other sample designs. For example, for a design without stratification, you can let $H = 1$; for a sample design without clustering, you can let $m_{hi} = 1$ for every h and i, which replaces clusters with individual sampling units.

For a two-way table representing the crosstabulation of two variables, define the following, where there are R levels of the row variable and C levels of the column variable:

r	$=$	$1, 2, \ldots, R$	is the row number, with a total of R rows
c	$=$	$1, 2, \ldots, C$	is the column number, with a total of C columns
N_{rc}			is the population total in row r and column c
$N_{r\cdot}$	$=$	$\sum_{c=1}^{C} N_{rc}$	is the total in row r
$N_{\cdot c}$	$=$	$\sum_{r=1}^{R} N_{rc}$	is the total in column c
N	$=$	$\sum_{r=1}^{R} \sum_{c=1}^{C} N_{rc}$	is the overall total

P_{rc}	$=$	$N_{rc} \;/\; N$	is the population proportion in row r and column c
$P_{r.}$	$=$	$N_{r\cdot} \;/\; N$	is the proportion in row r
$P_{.c}$	$=$	$N_{\cdot c} \;/\; N$	is the proportion in column c
$P_{rc}^{\,r}$	$=$	$N_{rc} \;/\; N_{r\cdot}$	is the row proportion for table cell (r, c)
$P_{rc}^{\,c}$	$=$	$N_{rc} \;/\; N_{\cdot c}$	is the column proportion for table cell (r, c)

For a specified observation (identified by stratum, cluster, and unit number within the cluster), define the following to indicate whether or not that observation belongs to cell (r, c), row r and column c, of the two-way table, for $r = 1, 2, \ldots, R$ and $c = 1, 2, \ldots, C$:

$$\delta_{hij}(r,c) = \begin{cases} 1 & \text{if observation } (hij) \text{ is in cell } (r,c) \\ 0 & \text{otherwise} \end{cases}$$

Similarly, define the following functions to indicate the observation's row and column classification:

$$\delta_{hij}(r\,\cdot) = \begin{cases} 1 & \text{if observation } (hij) \text{ is in row } r \\ 0 & \text{otherwise} \end{cases}$$

$$\delta_{hij}(\cdot\,c) = \begin{cases} 1 & \text{if observation } (hij) \text{ is in column } c \\ 0 & \text{otherwise} \end{cases}$$

Totals

PROC SURVEYFREQ estimates population frequency totals for the specified crosstabulation tables, including totals for two-way table cells, rows, columns, and overall totals. The procedure computes the estimate of the total frequency in table cell (r, c) as the weighted frequency sum,

$$\widehat{N}_{rc} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \delta_{hij}(r,c)\; W_{hij}$$

Similarly, PROC SURVEYFREQ computes estimates of row totals, column totals, and overall totals as

$$\widehat{N}_{r\cdot} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \delta_{hij}(r\,\cdot)\; W_{hij}$$

$$\widehat{N}_{\cdot c} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \delta_{hij}(\cdot\, c)\; W_{hij}$$

$$\widehat{N} \;= \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij}$$

PROC SURVEYFREQ estimates the variances of totals by using the variance estimation method that you request. If you request BRR variance estimation (by specifying the VARMETHOD=BRR option in the PROC SURVEYFREQ statement), the procedure estimates the variances as described in the section "Balanced Repeated Replication (BRR)" on page 8011. If you request jackknife variance estimation (by specifying the VARMETHOD=JACKKNIFE option), the procedure estimates the variances as described in the section "The Jackknife Method" on page 8014.

If you do not specify the VARMETHOD= option or a REPWEIGHTS statement, the default variance estimation method is Taylor series, which you can also request by specifying the VARMETHOD=TAYLOR option. Since totals are linear statistics, their variances can be estimated directly, without the approximation that is used for proportions and other nonlinear statistics. PROC SURVEYFREQ estimates the variance of the total frequency in table cell (r, c) as

$$\widehat{\mathrm{Var}}(\widehat{N}_{rc}) = \sum_{h=1}^{H} \widehat{\mathrm{Var}}_h(\widehat{N}_{rc})$$

where if $n_h > 1$,

$$\begin{aligned}
\widehat{\mathrm{Var}}_h(\widehat{N}_{rc}) &= \frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (n_{rc}^{hi} - \bar{n}_{rc}^{\,h})^2 \\
n_{rc}^{hi} &= \sum_{j=1}^{m_{hi}} \delta_{hij}(r,c)\; W_{hij} \\
\bar{n}_{rc}^{\,h} &= \left(\sum_{i=1}^{n_h} n_{rc}^{hi} \right) \Big/ \; n_h
\end{aligned}$$

and if $n_h = 1$,

$$\widehat{\mathrm{Var}}_h(\widehat{N}_{rc}) = \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1, 2, \ldots, H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 \le h' \le H \end{cases}$$

The standard deviation of the total is computed as

$$\mathrm{Std}(\widehat{N}_{rc}) = \sqrt{\widehat{\mathrm{Var}}(\widehat{N}_{rc})}$$

The variances and standard deviations are computed in a similar manner for row totals, column totals, and overall table totals.

Covariances of Frequency Estimates

The covariance matrix of the table cell frequency estimates is an $rc \times rc$ matrix that contains the pairwise cell frequency covariances. $\widehat{\mathbf{V}}(\widehat{\mathbf{N}})$ denotes the covariance matrix and $\widehat{\text{Cov}}(\widehat{N}_{rc}, \widehat{N}_{ab})$ denotes the pairwise covariances, for $r = 1, \ldots, R$; $c = 1, \ldots, C$; $a = 1, \ldots, R$; and $b = 1, \ldots, C$.

PROC SURVEYFREQ estimates the covariances by using the variance estimation method that you request. If you request BRR variance estimation (by specifying the VARMETHOD=BRR option in the PROC SURVEYFREQ statement), PROC SURVEYFREQ estimates the covariances by using the BRR method. If you request jackknife variance estimation (by specifying the VARMETHOD=JACKKNIFE option), PROC SURVEYFREQ estimates the covariances by using the jackknife method. For more information, see the sections “Balanced Repeated Replication (BRR)” on page 8011 and “The Jackknife Method” on page 8014.

By default, or if you request Taylor series variance estimation, PROC SURVEYFREQ estimates the covariance between frequency estimates for table cells (r,c) and (a,b) as

$$\widehat{\text{Cov}}(\widehat{N}_{rc}, \widehat{N}_{ab}) = \sum_{h=1}^{H} \left(\frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (n_{rc}^{hi} - \bar{n}_{rc}^{h})\ (n_{ab}^{hi} - \bar{n}_{ab}^{h}) \right)$$

Proportions

PROC SURVEYFREQ computes the estimate of the proportion in table cell (r, c) as the ratio of the estimated total for the table cell to the estimated overall total,

$$\begin{aligned} \widehat{P}_{rc} &= \widehat{N}_{rc} \ / \ \widehat{N} \\ &= \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \delta_{hij}(r,c)\ W_{hij} \right) \ / \ \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij} \right) \end{aligned}$$

If you request BRR variance estimation (by specifying the VARMETHOD=BRR option in the PROC SURVEYFREQ statement), the procedure estimates the variances of proportion estimates as described in the section “Balanced Repeated Replication (BRR)” on page 8011. If you request jackknife variance estimation (by specifying the VARMETHOD=JACKKNIFE option), the procedure estimates the variances as described in the section “The Jackknife Method” on page 8014.

If you do not specify the VARMETHOD= option or a REPWEIGHTS statement, the default variance estimation method is Taylor series, which you can also request by specifying the VARMETHOD=TAYLOR option. By using Taylor series linearization, the variance of a proportion estimate can be expressed as

$$\widehat{\text{Var}}(\widehat{P}_{rc}) = \sum_{h=1}^{H} \widehat{\text{Var}}_h(\widehat{P}_{rc})$$

where if $n_h > 1$,

$$\begin{aligned}
\widehat{\mathrm{Var}}_h(\widehat{P}_{rc}) &= \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h}(e_{rc}^{hi} - \bar{e}_{rc}^{h})^2 \\
e_{rc}^{hi} &= \left(\sum_{j=1}^{m_{hi}} (\delta_{hij}(r,c) - \widehat{P}_{rc})\; W_{hij} \right) / \; \widehat{N} \\
\bar{e}_{rc}^{h} &= \left(\sum_{i=1}^{n_h} e_{rc}^{hi} \right) / \; n_h
\end{aligned}$$

and if $n_h = 1$,

$$\widehat{\mathrm{Var}}_h(\widehat{P}_{rc}) = \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1, 2, \ldots, H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 \le h' \le H \end{cases}$$

The standard error of the proportion is computed as

$$\mathrm{StdErr}(\widehat{P}_{rc}) = \sqrt{\widehat{\mathrm{Var}}(\widehat{P}_{rc})}$$

Similarly, the estimate of the proportion in row r is

$$\widehat{P}_{r\cdot} = \widehat{N}_{r\cdot} \; / \; \widehat{N}$$

And its variance estimate is

$$\widehat{\mathrm{Var}}(\widehat{P}_{r\cdot}) = \sum_{h=1}^{H} \widehat{\mathrm{Var}}_h(\widehat{P}_{r\cdot})$$

where if $n_h > 1$,

$$\begin{aligned}
\widehat{\mathrm{Var}}_h(\widehat{P}_{r\cdot}) &= \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h}(e_{r\cdot}^{hi} - \bar{e}_{r\cdot}^{h})^2 \\
e_{r\cdot}^{hi} &= \left(\sum_{j=1}^{m_{hi}} (\delta_{hij}(r\,\cdot) - \widehat{P}_{r\cdot})\; W_{hij} \right) / \; \widehat{N} \\
\bar{e}_{r\cdot}^{h} &= \left(\sum_{i=1}^{n_h} e_{r\cdot}^{hi} \right) / \; n_h
\end{aligned}$$

and if $n_h = 1$,

$$\widehat{\mathrm{Var}}_h(\widehat{P}_{r\cdot}) = \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1, 2, \ldots, H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 \le h' \le H \end{cases}$$

The standard error of the proportion in row r is computed as

$$\mathrm{StdErr}(\widehat{P}_{r\cdot}) = \sqrt{\widehat{\mathrm{Var}}(\widehat{P}_{r\cdot})}$$

Computations for the proportion in column c are done in the same way.

Row and Column Proportions

PROC SURVEYFREQ computes the estimate of the row proportion for table cell (r, c) as the ratio of the estimated total for the table cell to the estimated total for row r,

$$\begin{aligned}
\widehat{P}^{\,r}_{rc} &= \widehat{N}_{rc} \;/\; \widehat{N}_{r\cdot} \\
&= \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \delta_{hij}(r,c)\; W_{hij} \right) \;/\; \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \delta_{hij}(r\;\cdot)\; W_{hij} \right)
\end{aligned}$$

Similarly, PROC SURVEYFREQ estimates the column proportion for table cell (r, c) as the ratio of the estimated total for the table cell to the estimated total for column c,

$$\begin{aligned}
\widehat{P}^{\,c}_{rc} &= \widehat{N}_{rc} \;/\; \widehat{N}_{\cdot c} \\
&= \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \delta_{hij}(r,c)\; W_{hij} \right) \;/\; \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \delta_{hij}(\cdot\; c)\; W_{hij} \right)
\end{aligned}$$

If you request BRR variance estimation (VARMETHOD=BRR), PROC SURVEYFREQ estimates the variances of the row and column proportions as described in the section "Balanced Repeated Replication (BRR)" on page 8011. If you request jackknife variance estimation (VARMETHOD=JACKKNIFE), the procedure estimates the variances as described in the section "The Jackknife Method" on page 8014.

If you do not specify the VARMETHOD= option or a REPWEIGHTS statement, the default variance estimation method is Taylor series (VARMETHOD=TAYLOR). By using Taylor series linearization, the variance of the row proportion estimate can be expressed as

$$\widehat{\mathrm{Var}}(\widehat{P}^{\,r}_{rc}) = \sum_{h=1}^{H} \widehat{\mathrm{Var}}_h(\widehat{P}_{rc})$$

where if $n_h > 1$,

$$\begin{aligned}
\widehat{\mathrm{Var}}_h(\widehat{P}^{\,r}_{rc}) &= \frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (g^{hi}_{rc} - \bar{g}^{\,h}_{rc})^2 \\
g^{hi}_{rc} &= \left(\sum_{j=1}^{m_{hi}} (\delta_{hij}(r,c) - \widehat{P}^{\,r}_{rc}\; \delta_{hij}(r\;\cdot))\; W_{hij} \right) \;/\; \widehat{N}_{r\cdot} \\
\bar{g}^{\,h}_{rc} &= \left(\sum_{i=1}^{n_h} g^{hi}_{rc} \right) \;/\; n_h
\end{aligned}$$

and if $n_h = 1$,

$$\widehat{\mathrm{Var}}_h(\widehat{P}^{\,r}_{rc}) = \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1, 2, \ldots, H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 \le h' \le H \end{cases}$$

The standard error of the row proportion is computed as

$$\mathrm{StdErr}(\widehat{P}_{rc}^{\,r}) = \sqrt{\widehat{\mathrm{Var}}(\widehat{P}_{rc}^{\,r})}$$

The Taylor series variance estimate for the column proportion is computed as described previously for the row proportion, but with

$$g_{rc}^{hi} = \left(\sum_{j=1}^{m_{hi}} (\delta_{hij}(r,c) - \widehat{P}_{rc}^{\,c}\, \delta_{hij}(\cdot\, c))\; W_{hij} \right) \,/\, \widehat{N}_{\cdot c}$$

Balanced Repeated Replication (BRR)

If you specify the VARMETHOD=BRR option, then PROC SURVEYFREQ uses balanced repeated replication (BRR) for variance estimation. The BRR variance estimation method requires a stratified sample design with two PSUs in each stratum. You can provide replicate weights for BRR variance estimation by using a REPWEIGHTS statement, or the procedure can construct replicate weights for the analysis. PROC SURVEYFREQ estimates the parameter of interest (a proportion, total, odds ratio, or other statistic) from each replicate, and then uses the variability among replicate estimates to estimate the overall variance of the parameter estimate. For more information about BRR variance estimation, see Wolter (1985) and Lohr (2010).

If you do not provide replicate weights with a REPWEIGHTS statement, PROC SURVEYFREQ constructs replicates based on the stratified design with two PSUs in each stratum. This section describes replicate construction by the traditional BRR method. If you specify the FAY *method-option* for VARMETHOD=BRR, the procedure uses Fay's modified BRR method, which is described in the section "Fay's BRR Method" on page 8012.

With the traditional BRR method, each replicate is obtained by deleting one PSU per stratum according to the corresponding Hadamard matrix of dimension R, where R is the number of replicates. The number of replicates equals the smallest multiple of 4 that is greater than the number of strata H. Alternatively, you can specify the number of replicates with the REPS= *method-option*. If a Hadamard matrix cannot be constructed for the REPS= value that you specify, the value is increased until a Hadamard matrix of that dimension can be constructed. Therefore, it is possible for the actual number of replicates used to be larger than the REPS= value that you specify.

You can provide a Hadamard matrix for BRR replicate construction by using the HADAMARD= *method-option*. Otherwise, PROC SURVEYFREQ generates an appropriate Hadamard matrix. For more information, see the section "Hadamard Matrix" on page 8013. You can display the Hadamard matrix by specifying the PRINTH *method-option*.

PROC SURVEYFREQ constructs replicates by using the first H columns of the $R \times R$ Hadamard matrix, where H denotes the number of strata. The rth replicate ($r = 1, 2, \ldots, R$) is drawn from the full sample according to the rth row of the Hadamard matrix as follows:

- If element (r, h) of the Hadamard matrix equals 1, then the first PSU of stratum h is included in the rth replicate, and the second PSU of stratum h is excluded.
- If element (r, h) of the Hadamard matrix equals –1, then the second PSU of stratum h is included in the rth replicate, and the first PSU of stratum h is excluded.

For the PSUs included in replicate r, the original weights are doubled to form the replicate r weights. For the PSUs not included in replicate r, the replicate r weights equal zero. You can use the OUTWEIGHTS=*SAS-data-set* *method-option* to store the replicate weights in a SAS data set. For information about the contents of the OUTWEIGHTS= data set, see the section "Replicate Weight Output Data Set" on page 8037. You can provide these replicate weights to the procedure for subsequent analyses by using a REPWEIGHTS statement.

Let θ denote the population parameter to be estimated—for example, a proportion, total, odds ratio, or other statistic. Let $\hat{\theta}$ denote the estimate of θ from the full sample, and let $\hat{\theta}_r$ denote the estimate from the rth BRR replicate, which is computed by using the replicate weights. The BRR variance estimate for $\hat{\theta}$ is computed as

$$\widehat{V}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}_r - \hat{\theta} \right)^2$$

where R is the total number of replicates.

If a parameter cannot be estimated from some replicate(s), then the variance estimate is computed by using those replicates from which the parameter can be estimated. For example, suppose the parameter is a column proportion—the proportion of column j for table cell (i, j). If a replicate r contains no observations in column j, then the column j proportion is not estimable from replicate r. In this case, the BRR variance estimate is computed as

$$\widehat{V}(\hat{\theta}) = \frac{1}{R'} \sum_{r=1}^{R'} \left(\hat{\theta}_r - \hat{\theta} \right)^2$$

where the summation is over the replicates where the parameter θ is estimable, and R' is the number of those replicates.

Fay's BRR Method

If you specify the FAY *method-option* for VARMETHOD=BRR, then PROC SURVEYFREQ uses Fay's BRR method, which is a modification of the traditional BRR variance estimation method. As for traditional BRR, Fay's method requires a stratified sample design with two PSUs in each stratum. You can provide replicate weights by using a REPWEIGHTS statement, or the procedure can construct replicate weights for the analysis. PROC SURVEYFREQ estimates the parameter of interest (a proportion, total, odds ratio, or other statistic) from each replicate, and then uses the variability among replicate estimates to estimate the overall variance of the parameter estimate.

If you do not provide replicate weights with a REPWEIGHTS statement, PROC SURVEYFREQ constructs replicates based on the stratified design with two PSUs in each stratum. As for traditional BRR, the number of replicates R equals the smallest multiple of 4 that is greater than the number of strata H, or you can specify the number of replicates with the REPS= *method-option*. You can provide a Hadamard matrix for replicate construction in the HADAMARD= *method-option*, or PROC SURVEYFREQ generates an appropriate Hadamard matrix.

The traditional BRR method constructs half-sample replicates by deleting one PSU per stratum according to the Hadamard matrix and doubling the original weights to form replicate weights. Fay's BRR method adjusts the original weights by a coefficient ϵ, where $0 \le \epsilon < 1$. You can specify the value of ϵ with the FAY= *method-option*. If you do not specify the value of ϵ, PROC SURVEYFREQ uses $\epsilon = 0.5$ by default. For information about the value of the Fay coefficient, see Judkins (1990) and Rao and Shao (1999). When

$\epsilon = 0$, Fay's method becomes the traditional BRR method. For more information, see Dippo, Fay, and Morganstein (1984), Fay (1989), and Judkins (1990).

PROC SURVEYFREQ constructs Fay BRR replicates by using the first H columns of the $R \times R$ Hadamard matrix, where H denotes the number of strata. The rth replicate ($r = 1, 2, \ldots, R$) is drawn from the full sample according to the rth row of the Hadamard matrix as follows:

- If element (r, h) of the Hadamard matrix equals 1, the sampling weight of the first PSU in stratum h is multiplied by ϵ, and the sampling weight of the second PSU is multiplied by $(2 - \epsilon)$ to form the rth replicate weights.
- If element (r, h) of the Hadamard matrix equals –1, then the sampling weight of the second PSU in stratum h is multiplied by ϵ, and the sampling weight of the first PSU is multiplied by $(2 - \epsilon)$ to form the rth replicate weights.

You can use the OUTWEIGHTS= *method-option* to store the replicate weights in a SAS data set. For information about the contents of the OUTWEIGHTS= data set, see the section "Replicate Weight Output Data Set" on page 8037. You can provide these replicate weights to the procedure for subsequent analyses by using a REPWEIGHTS statement.

Let θ denote the population parameter to be estimated—for example, a proportion, total, odds ratio, or other statistic. Let $\hat{\theta}$ denote the estimate of θ from the full sample, and let $\hat{\theta}_r$ denote the estimate from the rth BRR replicate, which is computed by using the replicate weights. The Fay BRR variance estimate for $\hat{\theta}$ is computed as

$$\widehat{V}(\hat{\theta}) = \frac{1}{R(1-\epsilon)^2} \sum_{r=1}^{R} \left(\hat{\theta}_r - \hat{\theta}\right)^2$$

where R is the total number of replicates and ϵ is the Fay coefficient.

If you request Fay's BRR method and also include a REPWEIGHTS statement, PROC SURVEYFREQ uses the replicate weights that you provide and includes the Fay coefficient ϵ in the denominator of the variance estimate in the preceding expression.

If a parameter cannot be estimated from some replicate(s), then the variance estimate is computed by using those replicates from which the parameter can be estimated. For example, suppose the parameter is a column proportion—the proportion of column j for table cell (i, j). If a replicate r contains no observations in column j, then the column j proportion is not estimable from replicate r. In this case, the BRR variance estimate is computed as

$$\widehat{V}(\hat{\theta}) = \frac{1}{R'(1-\epsilon)^2} \sum_{r=1}^{R'} \left(\hat{\theta}_r - \hat{\theta}\right)^2$$

where the summation is over the replicates where the parameter θ is estimable, and R' is the number of those replicates.

Hadamard Matrix

PROC SURVEYFREQ uses a Hadamard matrix to construct replicates for BRR variance estimation. You can provide a Hadamard matrix for replicate construction in the HADAMARD= *method-option* for

VARMETHOD=BRR. Otherwise, PROC SURVEYFREQ generates an appropriate Hadamard matrix. You can display the Hadamard matrix by specifying the PRINTH *method-option*.

A Hadamard matrix $\mathbf{A}$ of dimension R is a square matrix that has all elements equal to 1 or –1. A Hadamard matrix must satisfy the requirement that $\mathbf{A}'\mathbf{A} = R\mathbf{I}$, where $\mathbf{I}$ is an identity matrix. The dimension of a Hadamard matrix must equal 1, 2, or a multiple of 4.

For example, the following matrix is a Hadamard matrix of dimension $k = 8$:

$$\begin{array}{rrrrrrrr}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1
\end{array}$$

For BRR replicate construction, the dimension of the Hadamard matrix must be at least H, where H denotes the number of first-stage strata in your design. If a Hadamard matrix of a given dimension exists, it is not necessarily unique. Therefore, if you want to use a specific Hadamard matrix, you must provide the matrix as a SAS data set in the HADAMARD=*SAS-data-set method-option*. You must ensure that the matrix that you provide is actually a Hadamard matrix; PROC SURVEYFREQ does not check the validity of your Hadamard matrix.

For information about how the Hadamard matrix is used to construct replicates for BRR variance estimation, see the sections "Balanced Repeated Replication (BRR)" on page 8011 and "Fay's BRR Method" on page 8012.

The Jackknife Method

If you specify the VARMETHOD=JACKKNIFE option, PROC SURVEYFREQ uses the delete-1 jackknife method for variance estimation. The jackknife method can be used for stratified sample designs and for designs with no stratification. If your design is stratified, the jackknife method requires at least two PSUs in each stratum. You can provide replicate weights for jackknife variance estimation by using a REPWEIGHTS statement, or the procedure can construct replicate weights for the analysis. PROC SURVEYFREQ estimates the parameter of interest (a proportion, total, odds ratio, or other statistic) from each replicate, and then uses the variability among replicate estimates to estimate the overall variance of the parameter estimate. For more information about jackknife variance estimation, see Wolter (1985) and Lohr (2010).

If you do not provide replicate weights with a REPWEIGHTS statement, PROC SURVEYFREQ constructs the replicates. The number of replicates R equals the number of PSUs, and the procedure deletes one PSU from the full sample to form each replicate. The sampling weights are modified by the jackknife coefficient for the replicate to create the replicate weights.

If your design is not stratified (no STRATA statement), the jackknife coefficient has the same value for each replicate r. The jackknife coefficient equals

$$\alpha_r = (R-1)/R \quad \text{for } r = 1, 2, \ldots, R$$

where R is the total number of replicates (or total number of PSUs). For the PSUs included in a replicate, the replicate weights are computed by dividing the original sampling weights by the jackknife coefficient. For the deleted PSU, which is not included in the replicate, the replicate weights equal zero. The replicate weight for the *j*th member of the *i*th PSU can be expressed as follows when the design is not stratified:

$$W_{ij}^r = \begin{cases} W_{ij}/\alpha_r & \text{if PSU } i \text{ is included in replicate } r \\ 0 & \text{otherwise} \end{cases}$$

where W_{ij} is the original sampling weight of unit (ij), r is the replicate number, and α_r is the jackknife coefficient.

If your design is stratified, the jackknife method requires at least two PSUs in each stratum. Let stratum h_r' be the stratum from which a PSU is deleted to form the *r*th replicate. Stratum h_r' is called the *donor stratum*. The jackknife coefficients are defined as

$$\alpha_r = (n_{h_r'} - 1)/n_{h_r'} \quad \text{for } r = 1, 2, \ldots, R$$

where $n_{h_r'}$ is the total number of PSUs in the donor stratum for replicate *r*. For all strata other than the donor stratum, the replicate *r* weights equal the original sampling weights. For PSUs included from the donor stratum, the replicate weights are computed by dividing the original sampling weights by the jackknife coefficient. For the deleted PSU, which is not included in the replicate, the replicate weights equal zero. The replicate weight for the *j*th member of the *i*th PSU in stratum *h* can be expressed as

$$W_{hij}^r = \begin{cases} W_{hij} & \text{if } h \neq h_r' \\ W_{hij}/\alpha_r & \text{if } h = h_r' \text{ and PSU } (hi) \text{ is included in replicate } r \\ 0 & \text{if } h = h_r' \text{ and PSU } (hi) \text{ is not included in replicate } r \end{cases}$$

You can use the OUTWEIGHTS= *method-option* to store the replicate weights in a SAS data set. You can also use the OUTJKCOEFS= *method-option* to store the jackknife coefficients in a SAS data set. For information about the contents of these output data sets, see the sections “Jackknife Coefficient Output Data Set” on page 8038 and “Replicate Weight Output Data Set” on page 8037. You can provide replicate weights and jackknife coefficients to the procedure for subsequent analyses by using a REPWEIGHTS statement. If you provide replicate weights but do not provide jackknife coefficients, PROC SURVEYFREQ uses $\alpha_r = (R-1)/R$ as the jackknife coefficient for all replicates.

Let θ denote the population parameter to be estimated—for example, a proportion, total, odds ratio, or other statistic. Let $\hat{\theta}$ denote the estimate of θ from the full sample, and let $\hat{\theta}_r$ be the estimate from the *r*th jackknife replicate, which is computed by using the replicate weights. The jackknife variance estimate for $\hat{\theta}$ is computed as

$$\widehat{V}(\hat{\theta}) = \sum_{r=1}^{R} \alpha_r \left(\hat{\theta}_r - \hat{\theta}\right)^2$$

where R is the total number of replicates and α_r is the jackknife coefficient for replicate *r*.

If a parameter cannot be estimated from some replicate(s), then the variance estimate is computed by using those replicates from which the parameter can be estimated. For example, suppose the parameter is a column proportion—the proportion of column *j* for table cell (*i*, *j*). If a replicate *r* contains no observations in column *j*, then the column *j* proportion is not estimable from replicate *r*. In this case, the jackknife variance estimate is computed as

$$\widehat{V}(\hat{\theta}) = \frac{R}{R'} \sum_{r=1}^{R'} \alpha_r \left(\hat{\theta}_r - \hat{\theta}\right)^2$$

where the summation is over the replicates where the parameter θ is estimable, and R' is the number of those replicates.

Confidence Limits for Totals

If you specify the CLWT option in the TABLES statement, PROC SURVEYFREQ computes confidence limits for the weighted frequencies (totals) in the crosstabulation tables.

For the total in table cell (*r*, *c*), the confidence limits are computed as

$$\widehat{N}_{rc} \pm \left(t_{df,\alpha/2} \times \mathrm{StdErr}(\widehat{N}_{rc}) \right)$$

where $\widehat{N}_{rc}$ is the estimate of the total frequency in table cell (*r*, *c*), $\mathrm{StdErr}(\hat{N}_{rc})$ is the standard error of the estimate, and $t_{df,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the *t* distribution with *df* degrees of freedom. (For more information, see the section "Degrees of Freedom" on page 8019.) The confidence level α is determined by the value of the ALPHA= option; by default, ALPHA=0.05, which produces 95% confidence limits.

The confidence limits for row totals, column totals, and the overall total are computed similarly to the confidence limits for table cell totals.

For each table request, PROC SURVEYFREQ produces a nondisplayed ODS table, "Table Summary," which contains the number of observations, strata, and clusters that are included in the analysis of the requested table. When you request confidence limits, the "Table Summary" data set also contains the degrees of freedom *df* and the value of $t_{df,\alpha/2}$ that is used to compute the confidence limits. For more information about this output data set, see Example 97.3.

Confidence Limits for Proportions

If you specify the CL option in the TABLES statement, PROC SURVEYFREQ computes confidence limits for the proportions in the frequency and crosstabulation tables.

By default, PROC SURVEYFREQ computes Wald ("linear") confidence limits if you do not specify an alternative confidence limit type with the CL(TYPE=) option. In addition to Wald confidence limits, the following types of design-based confidence limits are available for proportions: modified Clopper-Pearson (exact), modified Wilson (score), and logit confidence limits.

PROC SURVEYFREQ also provides the CL(PSMALL) option, which uses the alternative confidence limit type for extreme (small or large) proportions and uses the Wald confidence limits for all other proportions (not extreme). For the default PSMALL= value of 0.25, the procedure computes Wald confidence limits for proportions between 0.25 and 0.75 and computes the alternative confidence limit type for proportions that are outside of this range. For more information, see Curtin et al. (2006).

For information about design-based confidence limits for proportions (including comparisons of their performance), see Korn and Graubard (1999), Korn and Graubard (1998), Curtin et al. (2006), and Sukasih and Jang (2005). For more information about binomial confidence limits, see Brown, Cai, and DasGupta (2001) and Agresti and Coull (1998), in addition to the references cited in the following sections.

For each table request, PROC SURVEYFREQ produces a nondisplayed ODS table, "Table Summary," which contains the number of observations, strata, and clusters that are included in the analysis of the requested table. When you request confidence limits, the "Table Summary" data set also contains the degrees of freedom *df* and the value of $t_{df,\alpha/2}$ that is used to compute the confidence limits. For more information about this output data set, see Example 97.3.

Wald Confidence Limits

PROC SURVEYFREQ computes standard Wald ("linear") confidence limits for proportions by default. These confidence limits use the variance estimates that are based on the sample design. For the proportion in table cell (r, c), the Wald confidence limits are computed as

$$\widehat{P}_{rc} \pm \left(t_{df,\alpha/2} \times \mathrm{StdErr}(\widehat{P}_{rc}) \right)$$

where $\widehat{P}_{rc}$ is the estimate of the proportion in table cell (r, c), $\mathrm{StdErr}(\widehat{P}_{rc})$ is the standard error of the estimate, and $t_{df,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the t distribution with df degrees of freedom. (For more information, see the section "Degrees of Freedom" on page 8019.) The confidence level α is determined by the value of the ALPHA= option; by default, ALPHA=0.05, which produces 95% confidence limits.

The confidence limits for row proportions and column proportions are computed similarly to the confidence limits for table cell proportions.

Modified Confidence Limits

PROC SURVEYFREQ uses the modification described in Korn and Graubard (1998) to compute design-based Clopper-Pearson (exact) and Wilson (score) confidence limits. This modification substitutes the degrees-of-freedom adjusted effective sample size for the original sample size in the confidence limit computations.

The effective sample size n_e is computed as

$$n_e \;=\; n \;/\; \mathrm{Deff}$$

where n is the original sample size (unweighted frequency) that corresponds to the total domain of the proportion estimate, and Deff is the design effect.

If the proportion is computed for a table cell of a two-way table, then the domain is the two-way table, and the sample size n is the frequency of the two-way table. If the proportion is a row proportion, which is based on a two-way table row, then the domain is the row, and the sample size n is the frequency of the row.

The design effect for an estimate is the ratio of the actual variance (estimated based on the sample design) to the variance of a simple random sample with the same number of observations. For more information, see the section "Design Effect" on page 8020.

If you do not specify the CL(ADJUST=NO) option, the procedure applies a degrees-of-freedom adjustment to the effective sample size to compute the modified sample size. If you specify CL(ADJUST=NO), the procedure does not apply the adjustment and uses the effective sample size n_e in the confidence limit computations.

The modified sample size n_e^* is computed by applying a degrees-of-freedom adjustment to the effective sample size n_e as

$$n_e^* \;=\; n_e \left(\frac{t_{(n-1),\alpha/2}}{t_{df,\alpha/2}} \right)^2$$

where df is the degrees of freedom and $t_{df,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the t distribution with df degrees of freedom. The degrees of freedom computation depends on the sample design and the variance estimation method. For more information, see the section "Degrees of Freedom" on page 8019. The confidence level α is determined by the value of the ALPHA= option; by default, ALPHA=0.05, which produces 95% confidence limits.

The design effect is usually greater than 1 for complex survey designs, and in that case the effective sample size is less than the actual sample size. If the adjusted effective sample size n_e^* is greater than the actual sample size n, the procedure truncates the value of n_e^* to n, as recommended by Korn and Graubard (1998). If you specify the CL(TRUNCATE=NO) option, the procedure does not truncate the value of n_e^*.

Modified Clopper-Pearson Confidence Limits Clopper-Pearson (exact) confidence limits for the binomial proportion are constructed by inverting the equal-tailed test based on the binomial distribution. This method is attributed to Clopper and Pearson (1934). For a derivation of the F distribution expression for the confidence limits, see Leemis and Trivedi (1996).

PROC SURVEYFREQ computes modified Clopper-Pearson confidence limits according to the approach of Korn and Graubard (1998). The degrees-of-freedom adjusted effective sample size n_e^* is substituted for the sample size in the Clopper-Pearson computation, and the adjusted effective sample size times the proportion estimate $n_e^* \hat{p}$ is substituted for the number of positive responses. (Or if you specify the CL(ADJUST=NO) option, the procedure uses the unadjusted effective sample size n_e instead of n_e^*.)

The modified Clopper-Pearson confidence limits for a proportion (P_L and P_U) are computed as

$$P_L = \left(1 + \frac{n_e^* - \hat{p} n_e^* + 1}{\hat{p} n_e^* \; F(\, \alpha/2, \; 2\hat{p} n_e^*, \; 2(n_e^* - \hat{p} n_e^* + 1)\,)} \right)^{-1}$$

$$P_U = \left(1 + \frac{n_e^* - \hat{p} n_e^*}{(\hat{p} n_e^* + 1) \; F(\, 1-\alpha/2, \; 2(\hat{p} n_e^* + 1), \; 2(n_e^* - \hat{p} n_e^*)\,)} \right)^{-1}$$

where $F(\alpha/2, b, c)$ is the $\alpha/2$ percentile of the F distribution with b and c degrees of freedom, n_e^* is the adjusted effective sample size, and $\hat{p}$ is the proportion estimate.

Modified Wilson Confidence Limits Wilson confidence limits for the binomial proportion are also known as score confidence limits and are attributed to Wilson (1927). The confidence limits are based on inverting the normal test that uses the null proportion in the variance (the score test). For more information, see Newcombe (1998) and Korn and Graubard (1999).

PROC SURVEYFREQ computes modified Wilson confidence limits by substituting the degrees-of-freedom adjusted effective sample size n_e^* for the original sample size in the standard Wilson computation. (Or if you specify the CL(ADJUST=NO) option, the procedure substitutes the unadjusted effective sample size n_e.)

The modified Wilson confidence limits for a proportion are computed as

$$\left(\hat{p} + (\kappa)^2 / 2n_e^* \right) \; \pm \; \left(\kappa \sqrt{ \left(\hat{p}(1-\hat{p}) + (\kappa)^2 \right) / 4n_e^* } \; / \; \left(1 + (\kappa)^2 / n_e^* \right) \right)$$

where n_e^* is the adjusted effective sample size and $\hat{p}$ is the estimate of the proportion. With the degrees-of-freedom adjusted effective sample size n_e^*, the computation uses $\kappa = z_{\alpha/2}$. With the unadjusted effective sample size, which you request with the ADJUST=NO option, the computation uses $\kappa = t_{df,\alpha/2}$. For more information, see Curtin et al. (2006).

Logit Confidence Limits

If you specify the CL(TYPE=LOGIT) option, PROC SURVEYFREQ computes logit confidence limits for proportions. For more information, see Agresti (2013) and Korn and Graubard (1998).

Logit confidence limits for proportions are based on the logit transformation $Y = \log(\hat{p}/(1-\hat{p}))$. The logit confidence limits P_L and P_U are computed as

$$P_L = \exp(Y_L) \;/\; (\, 1 + \exp(Y_L)\,)$$

$$P_U = \exp(Y_U) \;/\; (\, 1 + \exp(Y_U)\,)$$

where

$$(\, Y_L,\; Y_U \,) \;=\; \log(\hat{p}/(1-\hat{p})) \;\pm\; \left(t_{df,\alpha/2} \;\times\; \mathrm{StdErr}(\hat{p}) \;/\; (\hat{p}(1-\hat{p}))\, \right)$$

where $\hat{p}$ is the estimate of the proportion, $\mathrm{StdErr}(\hat{p})$ is the standard error of the estimate, and $t_{df,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the t distribution with *df* degrees of freedom. (For more information, see the section "Degrees of Freedom" on page 8019.) The confidence level α is determined by the value of the ALPHA= option; by default, ALPHA=0.05, which produces 95% confidence limits.

Degrees of Freedom

PROC SURVEYFREQ uses the degrees of freedom of the variance estimator to obtain the *t*-percentile for confidence limits for proportions, totals, and other statistics. The procedure also uses the degrees of freedom in computing the *F* statistics for the Rao-Scott and Wald chi-square tests.

PROC SURVEYFREQ computes the degrees of freedom based on the variance estimation method and the sample design. Alternatively, you can use the DF= option in the TABLES statement to specify the degrees of freedom.

For Taylor series variance estimation, PROC SURVEYFREQ calculates the degrees of freedom (*df*) as the number of clusters minus the number of strata. If there are no clusters, then *df* equals the number of observations minus the number of strata. If the design is not stratified, then *df* equals the number of clusters minus one. These numbers are based on the observations included in the analysis of the individual table request. These numbers do not count observations that are excluded from the table due to missing values. For more information, see the section "Missing Values" on page 8001. If you specify the MISSING option, missing values are treated as valid nonmissing levels and are included when computing degrees of freedom. If you specify the NOMCAR option for Taylor series variance estimation, observations with missing values of the TABLES variables are included when computing degrees of freedom.

If you use a REPWEIGHTS statement to provide replicate weights, the degrees of freedom equal the number of replicates, which is the number of REPWEIGHTS variables that you provide. Alternatively, you can use the DF= option in the REPWEIGHTS or the TABLES statement to specify the degrees of freedom.

For BRR variance estimation (when you do not use a REPWEIGHTS statement), PROC SURVEYFREQ calculates the degrees of freedom as the number of strata. PROC SURVEYFREQ bases the number of strata on all valid observations in the data set, unless you specify the DFADJ *method-option* for VARMETHOD=BRR. When you specify the DFADJ option, the procedure computes the degrees of freedom as the number of nonmissing strata for the individual table request. This excludes any empty strata that occur when observations with missing values of the TABLES variables are removed from the analysis for that table.

For jackknife variance estimation (when you do not use a REPWEIGHTS statement), PROC SURVEYFREQ calculates the degrees of freedom as the number of clusters minus the number of strata. If there are no clusters, then *df* equals the number of observations minus the number of strata. If the design is not stratified, then *df* equals the number of clusters minus one. For jackknife variance estimation, PROC SURVEYFREQ bases the number of strata and clusters on all valid observations in the data set, unless you specify the DFADJ *method-option* for VARMETHOD=JACKKNIFE. When you specify the DFADJ option, the procedure computes the degrees of freedom from the number of nonmissing strata and clusters for the individual table request. This excludes any empty strata or clusters that occur when observations with missing values of the TABLES variables are removed from the analysis for that table.

For each table request, PROC SURVEYFREQ produces a nondisplayed ODS table, "Table Summary," which contains the number of (nonmissing) observations, strata, and clusters that are included in the analysis of the table. If there are missing observations, empty strata, or empty clusters excluded from the analysis, the "Table Summary" data set also contains this information. If you request confidence limits or chi-square tests, which depend on the degrees of freedom of the variance estimator, the "Table Summary" data set provides the degrees of freedom *df*. For more information about this output data set, see Example 97.3.

Coefficient of Variation

If you specify the CV option in the TABLES statement, PROC SURVEYFREQ computes the coefficients of variation for the proportion estimates in the frequency and crosstabulation tables. The coefficient of variation is the ratio of the standard error to the estimate.

For the proportion in table cell (r, c), the coefficient of variation is computed as

$$\mathrm{CV}(\widehat{P}_{rc}) = \mathrm{StdErr}(\widehat{P}_{rc}) \;/\; \widehat{P}_{rc}$$

where $\widehat{P}_{rc}$ is the estimate of the proportion in table cell (r, c), and $\mathrm{StdErr}(\widehat{P}_{rc})$ is the standard error of the estimate. The coefficients of variation for row proportions and column proportions are computed similarly.

If you specify the CVWT option in the TABLES statement, PROC SURVEYFREQ computes the coefficients of variation for the weighted frequencies (estimated totals) in the crosstabulation tables. For the total in table cell (r, c), the coefficient of variation is computed as

$$\mathrm{CV}(\widehat{N}_{rc}) = \mathrm{StdErr}(\widehat{N}_{rc}) \;/\; \widehat{N}_{rc}$$

where $\widehat{N}_{rc}$ is the estimate of the total in table cell (r, c), and $\mathrm{StdErr}(\widehat{N}_{rc})$ is the standard error of the estimate. The coefficients of variation for row totals, column totals, and the overall total are computed similarly.

Design Effect

If you specify the DEFF option in the TABLES statement, PROC SURVEYFREQ computes design effects for the overall proportion estimates in the frequency and crosstabulation tables. If you specify the ROW(DEFF) or COLUMN(DEFF) option, the procedure provides design effects for the row or column proportion estimates, respectively. The design effect for an estimate is the ratio of the actual variance (estimated based on the sample design) to the variance of a simple random sample with the same number of observations. For more information, see Lohr (2010) and Kish (1965).

For Taylor series variance estimation, PROC SURVEYFREQ computes the design effect for the proportion in table cell (r, c) as

$$\begin{aligned}\mathrm{Deff}(\widehat{P}_{rc}) &= \widehat{\mathrm{Var}}(\widehat{P}_{rc}) \;/\; \widehat{\mathrm{Var}}_{\mathrm{srs}}(\widehat{P}_{rc}) \\ &= \widehat{\mathrm{Var}}(\widehat{P}_{rc}) \;/\; \Big((1-f)\; \widehat{P}_{rc}\; (1-\widehat{P}_{rc}) \;/\; (n-1) \Big)\end{aligned}$$

where $\widehat{P}_{rc}$ is the estimate of the proportion in table cell (r, c), $\widehat{\mathrm{Var}}(\widehat{P}_{rc})$ is the variance of the estimate, f is the overall sampling fraction, and n is the sample size (unweighted frequency) for the two-way table.

For Taylor series variance estimation, PROC SURVEYFREQ determines the value of f, the overall sampling fraction, based on the RATE= or TOTAL= option. If you do not specify either of these options, PROC SURVEYFREQ assumes the value of f is negligible and does not use a finite population correction in the analysis, as described in the section "Population Totals and Sampling Rates" on page 8000. If you specify RATE=*value*, PROC SURVEYFREQ uses this value as the overall sampling fraction f. If you specify TOTAL=*value*, PROC SURVEYFREQ computes f as the ratio of the number of PSUs in the sample to the specified total.

If you specify stratum sampling rates with the RATE=*SAS-data-set* option, then PROC SURVEYFREQ computes stratum totals based on these stratum sampling rates and the number of sample PSUs in each stratum. The procedure sums the stratum totals to form the overall total, and computes f as the ratio of the number of sample PSUs to the overall total. Alternatively, if you specify stratum totals with the TOTAL=*SAS-data-set* option, then PROC SURVEYFREQ sums these totals to compute the overall total. The overall sampling fraction f is then computed as the ratio of the number of sample PSUs to the overall total.

For BRR and jackknife variance estimation, PROC SURVEYFREQ computes the design effect for the proportion in table cell (r, c) as

$$\begin{aligned} \mathrm{Deff}(\widehat{P}_{rc}) &= \widehat{\mathrm{Var}}(\widehat{P}_{rc}) \;/\; \widehat{\mathrm{Var}}_{\mathrm{srs}}(\widehat{P}_{rc}) \\ &= \widehat{\mathrm{Var}}(\widehat{P}_{rc}) \;/\; \left(\widehat{P}_{rc}\ (1 - \widehat{P}_{rc}) \;/\; (n-1) \right) \end{aligned}$$

where $\widehat{P}_{rc}$ is the estimate of the proportion in table cell (r, c), $\widehat{\mathrm{Var}}(\widehat{P}_{rc})$ is the variance of the estimate, and n is the sample size (unweighted frequency) for the two-way table. This computation does not include the overall sampling fraction.

The procedure computes design effects similarly for proportions in one-way frequency tables, and also for row and column proportions in two-way tables. In these design effect computations, the value of n is the sample size (unweighted frequency) that corresponds to the total domain of the proportion estimate. For table cell proportions of a two-way table, the domain is the two-way table and the sample size n is the frequency of the two-way table. For row proportions, which are based on a two-way table row, the domain is the row and the sample size n is the frequency of the row.

Expected Weighted Frequency

If you specify the EXPECTED option in the TABLES statement, PROC SURVEYFREQ computes expected weighted frequencies for the table cells in two-way tables. The expected weighted frequencies are computed under the null hypothesis that the row and column variables are independent. The expected weighted frequency for table cell (r, c) equals

$$E_{rc} = \widehat{N}_{r\cdot}\ \widehat{N}_{\cdot c} \;/\; \widehat{N}$$

where $\widehat{N}_{r\cdot}$ is the estimated total for row r, $\widehat{N}_{\cdot c}$ is the estimated total for column c, and $\widehat{N}$ is the estimated overall total. Equivalently, the expected weighted frequency can be expressed as

$$E_{rc} = \widehat{P}_{r\cdot}\ \widehat{P}_{\cdot c}\ \widehat{N}$$

These expected values are used in the design-based chi-square tests of independence, as described in the sections "Rao-Scott Chi-Square Test" on page 8028 and "Wald Chi-Square Test" on page 8035.

Risks and Risk Difference

The RISK option provides estimates of risks (binomial proportions) and risk differences for 2×2 tables, together with their standard errors and confidence limits. Risk statistics include the row 1 risk, row 2 risk, overall risk, and risk difference. If you specify the RISK option, PROC SURVEYFREQ provides both column 1 and column 2 risks. You can request only column 1 (or only column 2) risks by specifying the RISK1 (or RISK2) option.

The column 1 risk for row 1 is the row 1 proportion for table cell (1,1). The column 1 risk estimate is computed as the ratio of the estimated total for table cell (1,1) to the estimated total for row 1,

$$\widehat{P}_{11}^{(1)} = \widehat{N}_{11} \ / \ \widehat{N}_{1\cdot}$$

where the total estimates are computed as described in the section "Totals" on page 8006. The column 1 risk for row 2 is the row 2 proportion for table cell (2,1), which is estimated as

$$\widehat{P}_{21}^{(2)} = \widehat{N}_{21} \ / \ \widehat{N}_{2\cdot}$$

The overall column 1 risk is the overall proportion in column 1, and its estimate is computed as

$$\widehat{P}_{\cdot 1} = \widehat{N}_{\cdot 1} \ / \ \widehat{N}$$

The column 2 risk estimates are computed similarly.

The row 1 and row 2 risks are the same as the row proportions for a 2×2 table, and their variances are computed as described in the section "Row and Column Proportions" on page 8010. The overall risk is the overall proportion in the column, and its variance computation is described in the section "Proportions" on page 8008. Confidence limits for the column 1 risk for row 1 are computed as

$$\widehat{P}_{11}^{(1)} \pm \left(t_{df,\alpha/2} \times \mathrm{StdErr}(\widehat{P}_{11}^{(1)}) \right)$$

where $\mathrm{StdErr}(\widehat{P}_{11}^{(1)})$ is the standard error of the risk estimate and $t_{df,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the t distribution with df degrees of freedom. (For more information, see the section "Degrees of Freedom" on page 8019.) The value of the confidence coefficient α is determined by the ALPHA= option; by default, ALPHA=0.05, which produces 95% confidence limits. Confidence limits for the other risks are computed similarly.

The risk difference is defined as the row 1 risk minus the row 2 risk. The estimate of the column 1 risk difference $\widehat{RD}_1$ is computed as

$$\begin{aligned} \widehat{RD}_1 &= \widehat{P}_{11}^{(1)} - \widehat{P}_{21}^{(2)} \\ &= \left(\widehat{N}_{11} \ / \ \widehat{N}_{1\cdot} \right) - \left(\widehat{N}_{21} \ / \ \widehat{N}_{2\cdot} \right) \end{aligned}$$

The column 2 risk difference is computed similarly.

PROC SURVEYFREQ estimates the variance of the risk difference by using the variance estimation method that you request. If you request BRR variance estimation (VARMETHOD=BRR), the procedure estimates the variance as described in the section "Balanced Repeated Replication (BRR)" on page 8011. If you request jackknife variance estimation (VARMETHOD=JACKKNIFE), the procedure estimates the variance as described in the section "The Jackknife Method" on page 8014.

If you do not specify the VARMETHOD= option or a REPWEIGHTS statement, the default variance estimation method is Taylor series (VARMETHOD=TAYLOR). By using Taylor series linearization, the variance estimate for the column 1 risk difference $\widehat{\mathrm{Var}}(\widehat{RD}_1)$ can be expressed as

$$\widehat{\mathrm{Var}}(\widehat{RD}_1) = \widehat{\mathbf{D}}\ \widehat{\mathbf{V}}(\widehat{\mathbf{X}})\ \widehat{\mathbf{D}}'$$

where $\widehat{\mathbf{V}}(\widehat{\mathbf{X}})$ is the covariance matrix of $\widehat{\mathbf{X}}$,

$$\widehat{\mathbf{X}} = \left(\widehat{N}_{11},\ \ \widehat{N}_{1\cdot},\ \ \widehat{N}_{21},\ \ \widehat{N}_{2\cdot} \right)$$

and $\widehat{\mathbf{D}}$ is an array that contains the partial derivatives of the risk difference with respect to the elements of $\widehat{\mathbf{X}}$,

$$\widehat{\mathbf{D}} = \left(1/\widehat{N}_{1\cdot},\ \ -\widehat{N}_{11}/\widehat{N}_{1\cdot}^2,\ \ -1/\widehat{N}_{2\cdot},\ \ -\widehat{N}_{21}/\widehat{N}_{2\cdot}^2 \right)$$

For more information, see Wolter (1985, pp. 239–242). The variance estimate for the column 2 risk difference is computed similarly.

The standard error of the column 1 risk difference is

$$\mathrm{StdErr}(\widehat{RD}_1) = \sqrt{\widehat{\mathrm{Var}}(\widehat{RD}_1)}$$

Confidence limits for the column 1 risk difference are computed as

$$\widehat{RD}_1 \pm \left(t_{df,\alpha/2} \times \mathrm{StdErr}(\widehat{RD}_1) \right)$$

where $t_{df,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the t distribution with df degrees of freedom. (For more information, see the section “Degrees of Freedom” on page 8019.) The value of the confidence coefficient α is determined by the ALPHA= option; by default, ALPHA=0.05, which produces 95% confidence limits. Confidence limits for the column 2 risk difference are computed in the same way.

Odds Ratio and Relative Risks

The OR option provides estimates of the odds ratio, the column 1 relative risk, and the column 2 relative risk for 2×2 tables, together with their confidence limits.

Odds Ratio

For a 2×2 table, the odds of a positive (column 1) response in row 1 is N_{11}/N_{12}. Similarly, the odds of a positive response in row 2 is N_{21}/N_{22}. The odds ratio is formed as the ratio of the row 1 odds to the row 2 odds. The estimate of the odds ratio is computed as

$$\widehat{OR} = \frac{\widehat{N}_{11}\ /\ \widehat{N}_{12}}{\widehat{N}_{21}\ /\ \widehat{N}_{22}} = \frac{\widehat{N}_{11}\ \widehat{N}_{22}}{\widehat{N}_{12}\ \widehat{N}_{21}}$$

The value of the odds ratio can be any nonnegative number. When the row and column variables are independent, the true value of the odds ratio equals 1. An odds ratio greater than 1 indicates that the odds of a positive response are higher in row 1 than in row 2. An odds ratio less than 1 indicates that the odds of positive response are higher in row 2. The strength of association increases with the deviation from 1. For more information, see Stokes, Davis, and Koch (2000) and Agresti (2007).

PROC SURVEYFREQ constructs confidence limits for the odds ratio by using the log transform. The $100(1-\alpha)\%$ confidence limits for the odds ratio are computed as

$$\left(\widehat{OR} \times \exp(-t_{df,\alpha/2}\ \sqrt{v}),\ \ \widehat{OR} \times \exp(t_{df,\alpha/2}\ \sqrt{v}) \right)$$

where

$$v = \widehat{\mathrm{Var}}(\ln \widehat{OR}) = \widehat{\mathrm{Var}}(\widehat{OR})\ /\ \widehat{OR}^{\,2}$$

is the estimate of the variance of the log odds ratio and $t_{df,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the t distribution with *df* degrees of freedom. (For more information, see the section "Degrees of Freedom" on page 8019.) The value of the confidence coefficient α is determined by the ALPHA= option; by default, ALPHA=0.05, which produces 95% confidence limits.

If you request BRR variance estimation (VARMETHOD=BRR), PROC SURVEYFREQ estimates the variance of the odds ratio as described in the section "Balanced Repeated Replication (BRR)" on page 8011. If you request jackknife variance estimation (VARMETHOD=JACKKNIFE), the procedure estimates the variance as described in the section "The Jackknife Method" on page 8014.

If you do not specify the VARMETHOD= option or a REPWEIGHTS statement, the default variance estimation method is Taylor series (VARMETHOD=TAYLOR). By using Taylor series linearization, the variance estimate for the odds ratio can be expressed as

$$\widehat{\mathrm{Var}}(\widehat{OR}) = \widehat{\mathbf{D}}\ \widehat{\mathbf{V}}(\widehat{\mathbf{N}})\ \widehat{\mathbf{D}}'$$

where $\widehat{\mathbf{V}}(\widehat{\mathbf{N}})$ is the covariance matrix of the estimates of the cell totals $\widehat{\mathbf{N}}$,

$$\widehat{\mathbf{N}} = \left(\widehat{N}_{11},\ \ \widehat{N}_{12},\ \ \widehat{N}_{21},\ \ \widehat{N}_{22} \right)$$

and $\widehat{\mathbf{D}}$ is an array that contains the partial derivatives of the odds ratio with respect to the elements of $\widehat{\mathbf{N}}$. The section "Covariances of Frequency Estimates" on page 8008 describes the computation of $\widehat{\mathbf{V}}(\widehat{\mathbf{N}})$. The array $\widehat{\mathbf{D}}$ is computed as

$$\widehat{\mathbf{D}} = \left(\widehat{N}_{22}/(\widehat{N}_{12}\widehat{N}_{21}),\quad -\widehat{N}_{11}\widehat{N}_{22}/(\widehat{N}_{21}\widehat{N}_{12}^{\,2}),\quad -\widehat{N}_{11}\widehat{N}_{22}/(\widehat{N}_{12}\widehat{N}_{21}^{\,2}),\quad \widehat{N}_{11}/(\widehat{N}_{12}\widehat{N}_{21}) \right)$$

For more information, see Wolter (1985, pp. 239–242).

Relative Risks

For a 2×2 table, the column 1 relative risk is the ratio of the column 1 risks for row 1 to row 2. As described in the section "Risks and Risk Difference" on page 8022, the column 1 risk for row 1 is the proportion of row 1 observations classified in column 1, and the column 1 risk for row 2 is the proportion of row 2 observations classified in column 1. The estimate of the column 1 relative risk is computed as

$$\widehat{RR}_1 = \frac{\widehat{N}_{11}\ /\ \widehat{N}_{1\cdot}}{\widehat{N}_{21}\ /\ \widehat{N}_{2\cdot}}$$

Similarly, the estimate of the column 2 relative risk is computed as

$$\widehat{RR}_2 = \frac{\widehat{N}_{12}\ /\ \widehat{N}_{1\cdot}}{\widehat{N}_{22}\ /\ \widehat{N}_{2\cdot}}$$

A relative risk greater than 1 indicates that the probability of positive response is greater in row 1 than in row 2. Similarly, a relative risk less than 1 indicates that the probability of positive response is less in row 1 than in row 2. The strength of association increases with the deviation from 1. For more information, see Stokes, Davis, and Koch (2000) and Agresti (2007).

PROC SURVEYFREQ constructs confidence limits for the relative risk by using the log transform, which is similar to the odds ratio computations described previously. The $100(1-\alpha)\%$ confidence limits for the column 1 relative risk are computed as

$$\left(\widehat{RR}_1 \times \exp(-t_{df,\alpha/2}\ \sqrt{v}),\ \widehat{RR}_1 \times \exp(t_{df,\alpha/2}\ \sqrt{v}) \right)$$

where

$$v = \widehat{\mathrm{Var}}(\ln \widehat{RR}_1) = \widehat{\mathrm{Var}}(\widehat{RR}_1)\ /\ \widehat{RR}_1^{\,2}$$

is the estimate of the variance of the log column 1 relative risk and $t_{df,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the t distribution with df degrees of freedom. (For more information, see the section "Degrees of Freedom" on page 8019.) The value of the confidence coefficient α is determined by the ALPHA= option; by default, ALPHA=0.05, which produces 95% confidence limits.

If you request BRR variance estimation (VARMETHOD=BRR), PROC SURVEYFREQ estimates the variance of the column 1 relative risk as described in the section "Balanced Repeated Replication (BRR)" on page 8011. If you request jackknife variance estimation (VARMETHOD=JACKKNIFE), the procedure estimates the variance as described in the section "The Jackknife Method" on page 8014.

If you do not specify the VARMETHOD= option or a REPWEIGHTS statement, the default variance estimation method is Taylor series (VARMETHOD=TAYLOR). By using Taylor series linearization, the variance estimate for the column 1 relative risk can be expressed as

$$\widehat{\mathrm{Var}}(\widehat{RR}_1) = \widehat{\mathbf{D}}\ \widehat{\mathbf{V}}(\widehat{\mathbf{X}})\ \widehat{\mathbf{D}}'$$

where $\widehat{\mathbf{V}}(\widehat{\mathbf{X}})$ is the covariance matrix of $\widehat{\mathbf{X}}$,

$$\widehat{\mathbf{X}} = \left(\widehat{N}_{11},\ \ \widehat{N}_{1\cdot},\ \ \widehat{N}_{21},\ \ \widehat{N}_{2\cdot} \right)$$

and $\widehat{\mathbf{D}}$ is an array that contains the partial derivatives of the column 1 relative risk with respect to the elements of $\widehat{\mathbf{X}}$,

$$\widehat{\mathbf{D}} = \left(\widehat{N}_{2\cdot}/(\widehat{N}_{21}\widehat{N}_{1\cdot}),\quad -\widehat{N}_{11}\widehat{N}_{2\cdot}/(\widehat{N}_{21}\widehat{N}_{1\cdot}^{2}),\quad -\widehat{N}_{11}\widehat{N}_{2\cdot}/(\widehat{N}_{1\cdot}\widehat{N}_{21}^{2}),\quad \widehat{N}_{11}/(\widehat{N}_{21}\widehat{N}_{1\cdot}) \right)$$

For more information, see Wolter (1985, pp. 239–242).

Confidence limits for the column 2 relative risk are computed similarly.

Kappa Coefficients

Simple Kappa Coefficient

The KAPPA option provides an estimate of the simple kappa coefficient, its standard error, and the confidence limits. This option is available with replication-based variance estimation methods (which you can request by specifying the VARMETHOD=JACKKNIFE or VARMETHOD=BRR option).

The simple kappa coefficient (Cohen 1960) is a measure of interrater agreement, where the row and column variables of the two-way table are viewed as two independent ratings. When there is perfect agreement between the two ratings, the kappa coefficient equals +1. When the observed agreement exceeds chance agreement, the value of kappa is positive, and its magnitude reflects the strength of agreement. The minimum value of kappa is between –1 and 0, depending on the marginal proportions. For more information, see Fleiss, Levin, and Paik (2003).

PROC SURVEYFREQ computes the simple kappa coefficient as

$$\hat{\kappa} = (P_o - P_e) \ / \ (1 - P_e)$$

where

$$P_o = \sum_i \widehat{P}_{ii}$$

$$P_e = \sum_i \left(\widehat{P}_{i\cdot} \widehat{P}_{\cdot i} \right)$$

where $\widehat{P}_{ii}$ is the estimate of the proportion in table cell (i, i), $\widehat{P}_{i\cdot}$ is the estimate of the proportion in row i, and $\widehat{P}_{\cdot i}$ is the estimate of the proportion in column i. For information about how PROC SURVEYFREQ computes the proportion estimates, see the section "Proportions" on page 8008.

If you request jackknife variance estimation (by specifying the VARMETHOD=JACKKNIFE option), PROC SURVEYFREQ estimates the variance of the simple kappa coefficient as described in the section "The Jackknife Method" on page 8014. If you request BRR variance estimation (by specifying the VARMETHOD=BRR option in the PROC SURVEYFREQ statement), the procedure estimates the variance as described in the section "Balanced Repeated Replication (BRR)" on page 8011.

PROC SURVEYFREQ computes confidence limits for the simple kappa coefficient as

$$\hat{\kappa} \pm \left(t_{df,\alpha/2} \times \mathrm{StdErr}(\hat{\kappa}) \right)$$

where $\mathrm{StdErr}(\hat{\kappa})$ is the standard error of the kappa coefficient and $t_{df,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the t distribution with df degrees of freedom. (For more information, see the section "Degrees of Freedom" on page 8019.) The value of the confidence coefficient α is determined by the ALPHA= option; by default, ALPHA=0.05, which produces 95% confidence limits.

Weighted Kappa Coefficient

The weighted kappa coefficient is a generalization of the simple kappa coefficient that uses agreement weights to quantify the relative difference between categories (levels). By default, PROC SURVEYFREQ uses Cicchetti-Allison agreement weights to compute the weighted kappa coefficient; if you specify the WTKAPPA(WT=FC) option, the procedure uses Fleiss-Cohen agreement weights. For information about how the agreement weights are computed, see the section "Kappa Agreement Weights" on page 8027. For more information, see Fleiss, Cohen, and Everitt (1969) and Fleiss, Levin, and Paik (2003).

For 2×2 tables, the weighted kappa coefficient equals the simple kappa coefficient; PROC SURVEYFREQ displays the weighted kappa coefficient only for tables larger than 2×2.

PROC SURVEYFREQ computes the weighted kappa coefficient as

$$\hat{\kappa}_w = \left(P_{o(w)} - P_{e(w)} \right) \ / \ \left(1 - P_{e(w)} \right)$$

where

$$P_{o(w)} = \sum_i \sum_j \left(w_{ij} \widehat{P}_{ij} \right)$$

$$P_{e(w)} = \sum_i \sum_j \left(w_{ij} \widehat{P}_{i\cdot} \widehat{P}_{\cdot j} \right)$$

where w_{ij} is the agreement weight for table cell (i, j), $\widehat{P}_{ij}$ is the estimate of the proportion in table cell (i, j), $\widehat{P}_{i\cdot}$ is the estimate of the proportion in row i, and $\widehat{P}_{\cdot i}$ is the estimate of the proportion in column i. For information about how PROC SURVEYFREQ computes the proportion estimates, see the section "Proportions" on page 8008.

If you request jackknife variance estimation (by specifying the VARMETHOD=JACKKNIFE option), PROC SURVEYFREQ estimates the variance of the weighted kappa coefficient as described in the section "The Jackknife Method" on page 8014. If you request BRR variance estimation (by specifying the VARMETHOD=BRR option in the PROC SURVEYFREQ statement), the procedure estimates the variance as described in the section "Balanced Repeated Replication (BRR)" on page 8011.

PROC SURVEYFREQ computes confidence limits for the weighted kappa coefficient as

$$\hat{\kappa}_w \pm \left(t_{df,\alpha/2} \times \mathrm{StdErr}(\hat{\kappa}_w) \right)$$

where $\mathrm{StdErr}(\hat{\kappa}_w)$ is the standard error of the weighted kappa coefficient and $t_{df,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the t distribution with df degrees of freedom. (For more information, see the section "Degrees of Freedom" on page 8019.) The value of the confidence coefficient α is determined by the ALPHA= option; by default, ALPHA=0.05, which produces 95% confidence limits.

Kappa Agreement Weights PROC SURVEYFREQ computes the weighted kappa coefficient by using the Cicchetti-Allison form (by default) or the Fleiss-Cohen form of agreement weights. These weights are based on the scores of the column variable in the two-way table request. If the column variable is numeric, the column scores are the numeric values of the column levels. If the column variable is a character variable, the column scores are the column numbers, where the columns are numbered in the order in which they appear in the crosstabulation table.

PROC SURVEYFREQ computes Cicchetti-Allison agreement weights as

$$w_{ij} = 1 - \left(|C_i - C_j| \;/\; (C_c - C_1) \right)$$

where C_i is the score for column i and c is the number of columns (categories). For more information, see Cicchetti and Allison (1971).

PROC SURVEYFREQ computes Fleiss-Cohen agreement weights as

$$w_{ij} = 1 - \left((C_i - C_j) \;/\; (C_c - C_1) \right)^2$$

For more information, see Fleiss and Cohen (1973).

The agreement weights w_{ij} are constructed so that $w_{ii} = 1$ for all i, and $w_{ij} = w_{ji}$. For $i \neq j$, the agreement weights must be nonnegative and less than 1, which is always true for character variables (where the scores are the column numbers). For numeric variables, you should assign numeric variable levels (scores) so that all agreement weights are nonnegative and less than 1.

You can assign numeric values to the variable levels in a way that reflects their degree of similarity. For example, suppose the column variable is numeric and has four levels, which you order according to similarity. If you assign the values 0, 2, 4, and 10 to the column variable levels, the Cicchetti-Allison agreement weights take the following values: $w_{12} = 0.8$, $w_{13} = 0.6$, $w_{14} = 0.0$, $w_{23} = 0.8$, $w_{24} = 0.2$, and $w_{34} = 0.4$. For this example, the Fleiss-Cohen agreement weights are as follows: $w_{12} = 0.96$, $w_{13} = 0.84$, $w_{14} = 0.00$, $w_{23} = 0.96$, $w_{24} = 0.36$, and $w_{34} = 0.64$.

To display the kappa agreement weights, you can specify the WTKAPPA(PRINTKWTS) option.

Rao-Scott Chi-Square Test

The Rao-Scott chi-square test is a design-adjusted version of the Pearson chi-square test, which involves differences between observed and expected frequencies. For information about design-adjusted chi-square tests, see Lohr (2010, Section 10.3.2), Rao and Scott (1981), Rao and Scott (1984), Rao and Scott (1987), and Thomas, Singh, and Roberts (1996).

PROC SURVEYFREQ provides a first-order Rao-Scott chi-square test by default. If you specify the CHISQ(SECONDORDER) option, PROC SURVEYFREQ provides a second-order (Satterthwaite) Rao-Scott chi-square test. The first-order design correction depends only on the design effects of the table cell proportion estimates and, for two-way tables, the design effects of the marginal proportion estimates. The second-order design correction requires computation of the full covariance matrix of the proportion estimates. The second-order test requires more computational resources than the first-order test, but it can provide some performance advantages (for Type I error and power), particularly when the design effects are variable (Thomas and Rao 1987; Rao and Thomas 1989).

One-Way Tables

For one-way tables, the CHISQ option provides a Rao-Scott (design-based) goodness-of-fit test for one-way tables. By default, this is a test for the null hypothesis of equal proportions. If you specify null hypothesis proportions in the TESTP= option, the goodness-of-fit test uses the specified proportions.

First-Order Test The first-order Rao-Scott chi-square statistic for the goodness-of-fit test is computed as

$$Q_{RS1} = Q_P \; / \; D$$

where Q_P is the Pearson chi-square based on the estimated totals and D is the first-order design correction described in the section "First-Order Design Correction" on page 8029. For more information, see Rao and Scott (1979), Rao and Scott (1981), Rao and Scott (1984).

For a one-way table with C levels, the Pearson chi-square is computed as

$$Q_P = (n / \widehat{N}) \; \sum_c (\widehat{N}_c - E_c)^2 \; / \; E_c$$

where n is the sample size, $\widehat{N}$ is the estimated overall total, $\widehat{N}_c$ is the estimated total for level c, and E_c is the expected total for level c under the null hypothesis. For the null hypothesis of equal proportions, the expected total for each level is

$$E_c = \widehat{N} \; / \; C$$

For specified null proportions, the expected total for level c equals

$$E_c = \widehat{N} \; \times \; {P_c}^0$$

where ${P_c}^0$ is the null proportion that you specify for level c.

Under the null hypothesis, the first-order Rao-Scott chi-square Q_{RS1} approximately follows a chi-square distribution with $(C - 1)$ degrees of freedom. A better approximation can be obtained by the F statistic,

$$F_1 = Q_{RS1} \; / \; (C - 1)$$

which has an F distribution with $(C - 1)$ and $\kappa(C - 1)$ degrees of freedom under the null hypothesis (Thomas and Rao 1984, 1987). The value of κ is the degrees of freedom for the variance estimator. The degrees of freedom computation depends on the sample design and the variance estimation method. For more information, see the section "Degrees of Freedom" on page 8019.

First-Order Design Correction By default for one-way tables, the first-order design correction is computed from the proportion estimates as

$$D = \sum_c (1 - \widehat{P}_c)\ \mathrm{Deff}(\widehat{P}_c)\ /\ (C - 1)$$

where

$$\begin{aligned} \mathrm{Deff}(\widehat{P}_c) &= \widehat{\mathrm{Var}}(\widehat{P}_c)\ /\ \mathrm{Var}_{\mathrm{srs}}(\widehat{P}_c) \\ &= \widehat{\mathrm{Var}}(\widehat{P}_c)\ /\ \Big((1 - f)\ \widehat{P}_c\ (1 - \widehat{P}_c)\ /\ (n - 1) \Big) \end{aligned}$$

as described in the section "Design Effect" on page 8020. $\widehat{P}_c$ is the proportion estimate for level c, $\widehat{\mathrm{Var}}(\widehat{P}_c)$ is the variance of the estimate, f is the overall sampling fraction, and n is the number of observations in the sample. The factor $(1 - f)$ is included only for Taylor series variance estimation (VARMETHOD=TAYLOR) when you specify the RATE= or TOTAL= option. For more information, see the section "Design Effect" on page 8020.

If you specify the CHISQ(MODIFIED) or LRCHISQ(MODIFIED) option, the design correction is computed by using null hypothesis proportions instead of proportion estimates. By default, null hypothesis proportions are equal proportions for all levels of the one-way table. Alternatively, you can specify null proportion values in the TESTP= option. The modified design correction D_0 is computed from null hypothesis proportions as

$$D_0 = \sum_c (1 - P_c{}^0)\ \mathrm{Deff}_0(\widehat{P}_c)\ /\ (C - 1)$$

where

$$\begin{aligned} \mathrm{Deff}_0(\widehat{P}_c) &= \widehat{\mathrm{Var}}(\widehat{P}_c)\ /\ \mathrm{Var}_{\mathrm{srs}}(P_c{}^0) \\ &= \widehat{\mathrm{Var}}(\widehat{P}_c)\ /\ \big((1 - f) P_c{}^0\ (1 - P_c{}^0)\ /\ (n - 1) \big) \end{aligned}$$

The null hypothesis proportion $P_c{}^0$ equals $1/C$ for equal proportions (the default), or $P_c{}^0$ equals the null proportion that you specify for level c if you use the TESTP= option.

Second-Order Test The second-order (Satterthwaite) Rao-Scott chi-square statistic for the goodness-of-fit test is computed as

$$Q_{RS2} = Q_{RS1}\ /\ (1 + \hat{a}^2)$$

where Q_{RS1} is the first-order Rao-Scott chi-square statistic described in the section "First-Order Test" on page 8028 and $\hat{a}^2$ is the second-order design correction described in the section "Second-Order Design Correction" on page 8030. For more information, see Rao and Scott (1979), Rao and Scott (1981), and Rao and Thomas (1989).

Under the null hypothesis, the second-order Rao-Scott chi-square Q_{RS2} approximately follows a chi-square distribution with $(C - 1)/(1 + \hat{a}^2)$ degrees of freedom. The corresponding F statistic is

$$F_{RS2} = Q_{RS2}\ /\ (C - 1)$$

which has an F distribution with $(C - 1)/(1 + \hat{a}^2)$ and $\kappa(C - 1)/(1 + \hat{a}^2)$ degrees of freedom under the null hypothesis (Thomas and Rao 1984, 1987). The value of κ is the degrees of freedom for the variance estimator. The degrees of freedom computation depends on the sample design and the variance estimation method. For more information, see the section "Degrees of Freedom" on page 8019.

Second-Order Design Correction The second-order (Satterthwaite) design correction for one-way tables is computed from the eigenvalues of the estimated design effects matrix $\widehat{\mathbf{\Delta}}$, which are known as *generalized design effects*. The design effects matrix is computed as

$$\widehat{\mathbf{\Delta}} \;=\; (n-1)/(1-f)\;\left(\mathbf{Cov_{srs}}(\widehat{\mathbf{P}})^{-1}\;\widehat{\mathbf{Cov}}(\widehat{\mathbf{P}})\right)$$

where $\mathbf{Cov_{srs}}(\widehat{\mathbf{P}})$ is the covariance under multinomial sampling (*srs* with replacement) and $\widehat{\mathbf{Cov}}(\widehat{\mathbf{P}})$ is the covariance matrix of the first $(C-1)$ proportion estimates. For more information, see Rao and Scott (1979), Rao and Scott (1981), and Rao and Thomas (1989).

By default, the *srs* covariance matrix is computed from the proportion estimates as

$$\mathbf{Cov_{srs}}(\widehat{\mathbf{P}}) \;=\; \mathrm{Diag}(\widehat{\mathbf{P}}) - \widehat{\mathbf{P}}\,\widehat{\mathbf{P}}\,'$$

where $\widehat{\mathbf{P}}$ is an array of $(C-1)$ proportion estimates. If you specify the CHISQ(MODIFIED) or LRCHISQ(MODIFIED) option, the *srs* covariance matrix is computed from the null hypothesis proportions $\mathbf{P_0}$ as

$$\mathbf{Cov_{srs}}(\mathbf{P_0}) \;=\; \mathrm{Diag}(\mathbf{P_0}) - \mathbf{P_0}\,\mathbf{P_0}'$$

where $\mathbf{P_0}$ is an array of $(C-1)$ null hypothesis proportions. The null hypothesis proportions equal $1/C$ by default. If you use the TESTP= option to specify null hypothesis proportions, $\mathbf{P_0}$ is an array of $(C-1)$ proportions that you specify.

The second-order design correction is computed as

$$\hat{a}^2 \;=\; \left(\sum_{c=1}^{C-1} d_c^2/(C-1)\bar{d}^2\right) - 1$$

where d_c are the eigenvalues of the design effects matrix $\widehat{\mathbf{\Delta}}$ and $\bar{d}$ is the average of the eigenvalues.

Two-Way Tables

For two-way tables, the CHISQ option provides a Rao-Scott (design-based) test of association between the row and column variables. PROC SURVEYFREQ provides a first-order Rao-Scott chi-square test by default. If you specify the CHISQ(SECONDORDER) option, PROC SURVEYFREQ provides a second-order (Satterthwaite) Rao-Scott chi-square test.

First-Order Test The first-order Rao-Scott chi-square statistic is computed as

$$Q_{RS1} = Q_P \;/\; D$$

where Q_P is the Pearson chi-square based on the estimated totals and D is the design correction described in the section "First-Order Design Correction" on page 8031. For more information, see Rao and Scott (1979), Rao and Scott (1984), and Rao and Scott (1987).

For a two-way tables with R rows and C columns, the Pearson chi-square is computed as

$$Q_P = (n/\widehat{N})\;\sum_r\sum_c(\widehat{N}_{rc} - E_{rc})^2 \;/\; E_{rc}$$

where n is the sample size, $\widehat{N}$ is the estimated overall total, $\widehat{N}_{rc}$ is the estimated total for table cell (r, c), and E_{rc} is the expected total for table cell (r,c) under the null hypothesis of no association,

$$E_{rc} = \widehat{N}_{r\cdot} \ \widehat{N}_{\cdot c} \ / \ \widehat{N}$$

Under the null hypothesis of no association, the first-order Rao-Scott chi-square Q_{RS1} approximately follows a chi-square distribution with $(R-1)(C-1)$ degrees of freedom. A better approximation can be obtained by the F statistic,

$$F_1 = Q_{RS1} \ / \ (R-1)(C-1)$$

which has an F distribution with $(R-1)(C-1)$ and $\kappa(R-1)(C-1)$ degrees of freedom under the null hypothesis (Thomas and Rao 1984, 1987). The value of κ is the degrees of freedom for the variance estimator. The degrees of freedom computation depends on the sample design and the variance estimation method. For more information, see the section “Degrees of Freedom” on page 8019.

First-Order Design Correction By default for a first-order test, PROC SURVEYFREQ computes the design correction from proportion estimates. If you specify the CHISQ(MODIFIED) or LRCHISQ(MODIFIED) option for a first-order test, the procedure computes the design correction from null hypothesis proportions.

Second-order tests, which you request by specifying the CHISQ(SECONDORDER) or LRCHISQ(SECONDORDER) option, are computed by applying both first-order and second-order design corrections to the weighted chi-square statistic. For second-order tests for two-way tables, PROC SURVEYFREQ always uses null hypothesis proportions to compute both the first-order and second-order design corrections.

The first-order design correction D that is based on proportion estimates is computed as

$$\begin{aligned} D = \ & \Big(\textstyle\sum_r \sum_c (1-\widehat{P}_{rc}) \, \mathrm{Deff}(\widehat{P}_{rc}) - \sum_r (1-\widehat{P}_{r\cdot}) \, \mathrm{Deff}(\widehat{P}_{r\cdot}) \\ & \quad - \textstyle\sum_c (1-\widehat{P}_{\cdot c}) \, \mathrm{Deff}(\widehat{P}_{\cdot c}) \Big) \ / \ (R-1)(C-1) \end{aligned}$$

where

$$\begin{aligned} \mathrm{Deff}(\widehat{P}_{rc}) &= \widehat{\mathrm{Var}}(\widehat{P}_{rc}) \ / \ \mathrm{Var}_{\mathrm{srs}}(\widehat{P}_{rc}) \\ &= \mathrm{Var}(\widehat{P}_{rc}) \ / \ \Big((1-f) \ \widehat{P}_{rc} \ (1-\widehat{P}_{rc}) \ / \ (n-1) \Big) \end{aligned}$$

as described in the section “Design Effect” on page 8020. $\widehat{P}_{rc}$ is the estimate of the proportion in table cell (r, c), $\widehat{\mathrm{Var}}(\widehat{P}_{rc})$ is the variance of the estimate, f is the overall sampling fraction, and n is the number of observations in the sample. The factor $(1-f)$ is included only for Taylor series variance estimation (VARMETHOD=TAYLOR) when you specify the RATE= or TOTAL= option. For more information, see the section “Design Effect” on page 8020.

The design effects for the estimate of the proportion in row r and the estimate of the proportion in column c ($\mathrm{Deff}(\widehat{P}_{r\cdot})$ and $\mathrm{Deff}(\widehat{P}_{\cdot c})$, respectively) are computed in the same way.

If you specify the CHISQ(MODIFIED) or LRCHISQ(MODIFIED) option for a first-order Rao-Scott test, or if you request a second-order test for a two-way table (CHISQ(SECONDORDER) or

LRCHISQ(SECONDORDER)), the procedure computes the design correction from the null hypothesis cell proportions instead of the estimated cell proportions. For two-way tables, the null hypothesis cell proportions are computed as the products of the corresponding row and column proportion estimates. The modified design correction D_0 (based on null hypothesis proportions) is computed as

$$\begin{aligned} D_0 = \; & \Big(\textstyle\sum_r \sum_c (1 - P_{rc}^0)\, \mathrm{Deff}_0(\widehat{P}_{rc}) - \sum_r (1 - \widehat{P}_{r\cdot})\, \mathrm{Deff}(\widehat{P}_{r\cdot}) \\ & \quad - \textstyle\sum_c (1 - \widehat{P}_{\cdot c})\, \mathrm{Deff}(\widehat{P}_{\cdot c}) \Big) \;/\; (R-1)(C-1) \end{aligned}$$

where

$$P_{rc}^0 = \widehat{P}_{r\cdot} \times \widehat{P}_{\cdot c}$$

and

$$\begin{aligned} \mathrm{Deff}_0(\widehat{P}_{rc}) &= \widehat{\mathrm{Var}}(\widehat{P}_{rc}) \;/\; \mathrm{Var}_{\mathrm{srs}}(P_{rc}^0) \\ &= \widehat{\mathrm{Var}}(\widehat{P}_{rc}) \;/\; \big((1-f)\; P_{rc}^0\; (1 - P_{rc}^0) \;/\; (n-1)\big) \end{aligned}$$

Second-Order Test The second-order (Satterthwaite) Rao-Scott chi-square statistic for two-way tables is computed as

$$Q_{RS2} = Q_{RS1} \;/\; (1 + \hat{a}^2)$$

where Q_{RS1} is the first-order Rao-Scott chi-square statistic described in the section "First-Order Test" on page 8030 and $\hat{a}^2$ is the second-order design correction described in the section "Second-Order Design Correction" on page 8032. For more information, see Rao and Scott (1979), Rao and Scott (1981), and Rao and Thomas (1989).

Under the null hypothesis, the second-order Rao-Scott chi-square Q_{RS2} approximately follows a chi-square distribution with $(R-1)(C-1)/(1+\hat{a}^2)$ degrees of freedom. The corresponding F statistic is

$$F_{RS2} = Q_{RS2}\; (1 + \hat{a}^2) \;/\; (R-1)(C-1)$$

which has an F distribution with $(R-1)(C-1)/(1+\hat{a}^2)$ and $\kappa(R-1)(C-1)/(1+\hat{a}^2)$ degrees of freedom under the null hypothesis (Thomas and Rao 1984, 1987). The value of κ is the degrees of freedom for the variance estimator. The degrees of freedom computation depends on the sample design and the variance estimation method. For more information, see the section "Degrees of Freedom" on page 8019.

Second-Order Design Correction The second-order (Satterthwaite) design correction for two-way tables is computed from the eigenvalues of the estimated design effects matrix $\widehat{\mathbf{\Delta}}$, which are known as *generalized design effects*. The design effects matrix is defined as

$$\widehat{\mathbf{\Delta}} \;=\; (n-1)/(1-f)\; \Big(\mathbf{Cov_{srs}}(\widehat{\mathbf{P}})^{-1}\; \mathbf{H}\; \widehat{\mathbf{Cov}}(\widehat{\mathbf{P}})\; \mathbf{H}' \Big)$$

where $\widehat{\mathbf{Cov}}(\widehat{\mathbf{P}})$ is the covariance matrix of the $R \times C$ proportion estimates and $\mathbf{Cov_{srs}}(\widehat{\mathbf{P}})$ is the covariance under multinomial sampling (*srs* with replacement). For more information, see Rao and Scott (1979), Rao and Scott (1981), and Rao and Thomas (1989).

The second-order design correction is computed from the design effects matrix $\widehat{\boldsymbol{\Delta}}$ as

$$\hat{a}^2 \;=\; \left(\sum_{i=1}^{K} d_c^2 / K \bar{d}^2 \right) - 1$$

where $K = (R-1)(C-1)$, d_c are the eigenvalues of $\widehat{\boldsymbol{\Delta}}$, and $\bar{d}$ is the average eigenvalue.

The *srs* covariance matrix is computed as

$$\mathbf{Cov_{srs}}(\widehat{\mathbf{P}}) \;=\; \widehat{\mathbf{P}}_\mathbf{r} \otimes \widehat{\mathbf{P}}_\mathbf{c}$$

where $\widehat{\mathbf{P}}_\mathbf{r}$ is an $(R-1) \times (R-1)$ matrix that is constructed from the array of $(R-1)$ row proportion estimates $\widehat{\mathbf{p}}_\mathbf{r}$ as

$$\widehat{\mathbf{P}}_\mathbf{r} \;=\; \mathrm{Diag}(\widehat{\mathbf{p}}_\mathbf{r}) - \widehat{\mathbf{p}}_\mathbf{r}\, \widehat{\mathbf{p}}_\mathbf{r}'$$

Similarly, $\widehat{\mathbf{P}}_\mathbf{c}$ is a $(C-1) \times (C-1)$ matrix that is constructed from the array of $(C-1)$ column proportion estimates $\widehat{\mathbf{p}}_\mathbf{c}$ as

$$\widehat{\mathbf{P}}_\mathbf{c} \;=\; \mathrm{Diag}(\widehat{\mathbf{p}}_\mathbf{c}) - \widehat{\mathbf{p}}_\mathbf{c}\, \widehat{\mathbf{p}}_\mathbf{c}'$$

The $(R-1)(C-1) \times (R-1)(C-1)$ matrix $\mathbf{H}$ is computed as

$$\mathbf{H} \;=\; \mathbf{J_r} \otimes \mathbf{J_c} \;-\; (\widehat{\mathbf{p}}_\mathbf{r}\, \mathbf{l}_\mathbf{r}') \otimes \mathbf{J_c} \;-\; \mathbf{J_r} \otimes (\widehat{\mathbf{p}}_\mathbf{r}\, \mathbf{l}_\mathbf{r}')$$

where $\mathbf{J_r} = (\mathbf{I}_{(R-1)}|\mathbf{0})$, $\mathbf{J_c} = (\mathbf{I}_{(C-1)}|\mathbf{0})$, $\mathbf{l_r}$ is an $(R \times 1)$ array of ones, and $\mathbf{l_c}$ is an $(C \times 1)$ array of ones. For more information, see Rao and Scott (1979, p. 61).

Rao-Scott Likelihood Ratio Chi-Square Test

The Rao-Scott likelihood ratio chi-square test is a design-adjusted version of the likelihood ratio test, which involves ratios of observed and expected frequencies. For information about design-adjusted chi-square tests, see Lohr (2010, Section 10.3.2), Rao and Scott (1981), Rao and Scott (1984), Rao and Scott (1987), and Thomas, Singh, and Roberts (1996).

PROC SURVEYFREQ provides a first-order Rao-Scott likelihood ratio test by default. If you specify the LRCHISQ(SECONDORDER) option, PROC SURVEYFREQ provides a second-order (Satterthwaite) likelihood ratio chi-square test.

The procedure computes the Rao-Scott likelihood ratio test by applying design adjustments to the weighted likelihood ratio statistic that is based on estimated totals. This computation is identical to the Rao-Scott chi-square test computation except that it uses the likelihood ratio statistic G^2 in place of the Pearson chi-square statistic Q_P. For more information, see the section "Rao-Scott Chi-Square Test" on page 8028.

One-Way Tables

For one-way tables, the LRCHISQ option provides a Rao-Scott (design-based) goodness-of-fit test for one-way tables. By default, this is a test for the null hypothesis of equal proportions. If you specify null hypothesis proportions in the TESTP= option, the goodness-of-fit test uses the specified proportions.

The Rao-Scott likelihood ratio test uses the likelihood ratio statistic that is based on the estimated totals,

$$G^2 = 2 \;\left(n \;/\; \widehat{N} \;\right)\; \sum_c \; \widehat{N}_c \;\; \ln \left(\widehat{N}_c \;/\; E_c \right)$$

where n is the sample size, $\widehat{N}$ is the estimated overall total, $\widehat{N}_c$ is the estimated total for level c, and E_c is the expected total for level c under the null hypothesis. For the null hypothesis of equal proportions, the expected total for each level equals

$$E_c = \widehat{N} \; / \; C$$

For specified null proportions, the expected total for level c equals

$$E_c = \widehat{N} \times {P_c}^0$$

where ${P_c}^0$ is the null proportion that you specify for level c.

The computation of the Rao-Scott likelihood ratio test for one-way tables uses G^2 in place of Q_P in the Rao-Scott chi-square test computation and is otherwise identical to the chi-square test computation. For more information, see the sections "First-Order Test" on page 8028 and "Second-Order Test" on page 8029.

If you specify the LRCHISQ(MODIFIED) option, PROC SURVEYFREQ computes the design corrections by using null hypothesis proportions instead of proportion estimates. By default, null hypothesis proportions are equal proportions for all levels of the one-way table. Alternatively, you can specify null proportion values in the TESTP= option.

Two-Way Tables

For two-way tables, the LRCHISQ option provides a Rao-Scott (design-based) test of association between the row and column variables.

The Rao-Scott likelihood ratio test uses the likelihood ratio statistic that is based on the estimated totals,

$$G^2 = 2 \; (n/\widehat{N}) \; \sum_r \sum_c \widehat{N}_{rc} \; \ln\left(\widehat{N}_{rc} \; / \; E_{rc}\right)$$

where n is the sample size, $\widehat{N}$ is the estimated overall total, $\widehat{N}_{rc}$ is the estimated total for table cell (r, c), and E_{rc} is the expected total for cell (r, c) under the null hypothesis of no association. The expected total for cell (r, c) equals

$$E_{rc} = \widehat{N}_{r\cdot} \; \widehat{N}_{\cdot c} \; / \; \widehat{N}$$

The computation of the Rao-Scott likelihood ratio test for two-way tables uses G^2 in place of Q_P in the Rao-Scott chi-square test computation and is otherwise identical to the chi-square test computation. For more information, see the sections "First-Order Test" on page 8030 and "Second-Order Test" on page 8032.

By default for a first-order test, PROC SURVEYFREQ computes the design correction from proportion estimates. If you specify the LRCHISQ(MODIFIED) option for a first-order test, the procedure computes the design correction from null hypothesis proportions.

Second-order tests, which you request by specifying the LRCHISQ(SECONDORDER) option, are computed by applying both first-order and second-order design corrections to the weighted likelihood ratio statistic. For second-order tests for two-way tables, PROC SURVEYFREQ always uses null hypothesis proportions to compute both the first-order and second-order design corrections.

Wald Chi-Square Test

PROC SURVEYFREQ provides two Wald chi-square tests for independence of the row and column variables in a two-way table: a Wald chi-square test based on the difference between observed and expected weighted cell frequencies, and a Wald log-linear chi-square test based on the log odds ratios. These statistics test for independence of the row and column variables in two-way tables, taking into account the complex survey design. For information about Wald statistics and their applications to categorical data analysis, see Bedrick (1983), Koch, Freeman, and Freeman (1975), and Wald (1943).

For these two tests, PROC SURVEYFREQ computes the generalized Wald chi-square statistic, the corresponding F statistic, and also an adjusted F statistic for tables larger than 2×2. Under the null hypothesis of independence, the Wald chi-square statistic approximately follows a chi-square distribution with $(R-1)(C-1)$ degrees of freedom for large samples. However, it has been shown that this test can perform poorly in terms of actual significance level and power, especially for tables with a large number of cells or for samples with a relatively small number of clusters. For more information, see Thomas and Rao (1984), Thomas and Rao (1985), and Lohr (2010). For information about the adjusted F statistic, see Felligi (1980) and Hidiroglou, Fuller, and Hickman (1980). Thomas and Rao (1984) found that the adjusted F statistic provides a more stable test than the chi-square statistic, although its power can be low when the number of sample clusters is not large. See also Korn and Graubard (1990) and Thomas, Singh, and Roberts (1996).

If you specify the WCHISQ option in the TABLES statement, PROC SURVEYFREQ computes a Wald test for independence in the two-way table based on the differences between the observed (weighted) cell frequencies and the expected frequencies.

Under the null hypothesis of independence of the row and column variables, the expected cell frequencies are computed as

$$E_{rc} = \widehat{N}_{r\cdot}\ \widehat{N}_{\cdot c}\ /\ \widehat{N}$$

where $\widehat{N}_{r\cdot}$ is the estimated total for row r, $\widehat{N}_{\cdot c}$ is the estimated total for column c, and $\widehat{N}$ is the estimated overall total, as described in the section "Expected Weighted Frequency" on page 8021. The null hypothesis that the population weighted frequencies equal the expected frequencies can be expressed as

$$H_0\colon Y_{rc} = N_{rc} - E_{rc} = 0$$

for all $r = 1, \ldots (R-1)$ and $c = 1, \ldots (C-1)$. This null hypothesis can be stated equivalently in terms of cell proportions, with the expected cell proportions computed as the products of the marginal row and column proportions.

The generalized Wald chi-square statistic Q_W is computed as

$$Q_W = \widehat{\mathbf{Y}}'\ (\mathbf{H}\ \widehat{\mathbf{V}}(\widehat{\mathbf{N}})\ \mathbf{H}')^{-1}\ \widehat{\mathbf{Y}}$$

where $\widehat{\mathbf{Y}}$ is an array of $(R-1)(C-1)$ differences between the observed and expected weighted frequencies $(\widehat{N}_{rc} - E_{rc})$, and $(\mathbf{H}\ \widehat{\mathbf{V}}(\widehat{\mathbf{N}})\ \mathbf{H}')$ estimates the variance of $\widehat{\mathbf{Y}}$.

$\widehat{\mathbf{V}}(\widehat{\mathbf{N}})$ is the covariance matrix of the estimates $\widehat{N}_{rc}$, and its computation is described in the section "Covariances of Frequency Estimates" on page 8008.

$\mathbf{H}$ is an $(R-1)(C-1)$ by RC matrix that contains the partial derivatives of the elements of $\widehat{\mathbf{Y}}$ with respect to the elements of $\widehat{\mathbf{N}}$. The elements of $\mathbf{H}$ are computed as follows, where a denotes a row different from row r, and b denotes a column different from column c:

$$\begin{aligned}
\partial \widehat{Y}_{rc} / \partial \widehat{N}_{rc} &= 1 - \left(\widehat{N}_{r\cdot} + \widehat{N}_{\cdot c} - \widehat{N}_{\cdot c}\, \widehat{N}_{r\cdot} \,/\, \widehat{N} \right) / \, \widehat{N} \\
\partial \widehat{Y}_{rc} / \partial \widehat{N}_{ac} &= - \left(\widehat{N}_{r\cdot} - \widehat{N}_{r\cdot}\, \widehat{N}_{\cdot c} \,/\, \widehat{N} \right) / \, \widehat{N} \\
\partial \widehat{Y}_{rc} / \partial \widehat{N}_{rb} &= - \left(\widehat{N}_{\cdot c} - \widehat{N}_{r\cdot}\, \widehat{N}_{\cdot c} \,/\, \widehat{N} \right) / \, \widehat{N} \\
\partial \widehat{Y}_{rc} / \partial \widehat{Y}_{ab} &= \widehat{N}_{r\cdot}\, \widehat{N}_{\cdot c} \,/\, \widehat{N}^{\,2}
\end{aligned}$$

Under the null hypothesis of independence, the statistic Q_W approximately follows a chi-square distribution with (R – 1)(C – 1) degrees of freedom for large samples.

PROC SURVEYFREQ computes the Wald F statistic as

$$F_W = Q_W \;/\; (R-1)(C-1)$$

Under the null hypothesis of independence, F_W approximately follows an F distribution with (R – 1)(C – 1) numerator degrees of freedom. The denominator degrees of freedom are the degrees of freedom for the variance estimator and depend on the sample design and the variance estimation method. For more information, see the section “Degrees of Freedom” on page 8019. Alternatively, you can use the DF= option in the TABLES statement to specify the denominator degrees of freedom.

For tables larger than 2 × 2, PROC SURVEYFREQ also computes the adjusted Wald F statistic as

$$F_{Adj\_W} = Q_W \; (s-k+1) \;/\; (ks)$$

where k = (R – 1)(C – 1), and s is the degrees of freedom. (For more information, see the section “Degrees of Freedom” on page 8019.) Alternatively, you can use the DF= option in the TABLES statement to specify the value of s. For 2 × 2 tables, k = (R – 1)(C – 1) = 1, and therefore the adjusted Wald F statistic equals the (unadjusted) Wald F statistic and has the same numerator and denominator degrees of freedom.

Under the null hypothesis, F_{Adj_W} approximately follows an F distribution with k numerator degrees of freedom and (s – k + 1) denominator degrees of freedom.

Wald Log-Linear Chi-Square Test

If you specify the WLLCHISQ option in the TABLES statement, PROC SURVEYFREQ computes a Wald test for independence based on the log odds ratios. For more information about Wald tests, see the section “Wald Chi-Square Test” on page 8035.

For a two-way table of R rows and C columns, the Wald log-linear test is based on the (R – 1)(C – 1)-dimensional array of elements $\widehat{Y}_{rc}$,

$$\widehat{Y}_{rc} = \log \widehat{N}_{rc} \;-\; \log \widehat{N}_{rC} \;-\; \log \widehat{N}_{Rc} \;+\; \log \widehat{N}_{RC}$$

where $\widehat{N}_{rc}$ is the estimated total for table cell (r, c). The null hypothesis of independence between the row and column variables can be expressed as $H_0\colon Y_{rc} = 0$ for all $r = 1, \ldots (R-1)$ and $c = 1, \ldots (C-1)$. This null hypothesis can be stated equivalently in terms of cell proportions.

The generalized Wald log-linear chi-square statistic is computed as

$$Q_L = \widehat{\mathbf{Y}}' \; \widehat{\mathbf{V}}(\widehat{\mathbf{Y}})^{-1} \; \widehat{\mathbf{Y}}$$

where $\widehat{\mathbf{Y}}$ is the $(R-1)(C-1)$-dimensional array of the $\widehat{Y}_{rc}$, and $\widehat{\mathbf{V}}(\widehat{\mathbf{Y}})$ estimates the variance of $\widehat{\mathbf{Y}}$,

$$\widehat{\mathbf{V}}(\widehat{\mathbf{Y}}) = \mathbf{A}\ \mathbf{D}^{-1}\ \widehat{V}(\widehat{\mathbf{N}})\ \mathbf{D}^{-1}\ \mathbf{A}'$$

where $\widehat{\mathbf{V}}(\widehat{\mathbf{N}})$ is the covariance matrix of the estimates $\widehat{N}_{rc}$, which is computed as described in the section "Covariances of Frequency Estimates" on page 8008. $\mathbf{D}$ is a diagonal matrix with the estimated totals $\widehat{N}_{rc}$ on the diagonal, and $\mathbf{A}$ is the $(R-1)(C-1)$ by $RC \times RC$ linear contrast matrix.

Under the null hypothesis of independence, the statistic Q_L approximately follows a chi-square distribution with $(R-1)(C-1)$ degrees of freedom for large samples.

PROC SURVEYFREQ computes the Wald log-linear F statistic as

$$F_L = Q_L \;/\; (R-1)(C-1)$$

Under the null hypothesis of independence, F_L approximately follows an F distribution with $(R-1)(C-1)$ numerator degrees of freedom. PROC SURVEYFREQ computes the denominator degrees of freedom as described in the section "Degrees of Freedom" on page 8019. Alternatively, you can use the DF= option in the TABLES statement to specify the denominator degrees of freedom.

For tables larger than 2×2, PROC SURVEYFREQ also computes the adjusted Wald log-linear F statistic as

$$F_{Adj\_L} = Q_L\ (s-k+1)\ /\ (ks)$$

where $k = (R-1)(C-1)$, and s is the denominator degrees of freedom, which is computed as described in the section "Degrees of Freedom" on page 8019. Alternatively, you can use the DF= option in the TABLES statement to specify the value of s. For 2×2 tables, $k = (R-1)(C-1) = 1$, and therefore the adjusted Wald F statistic equals the (unadjusted) Wald F statistic and has the same numerator and denominator degrees of freedom.

Under the null hypothesis, F_{Adj_L} approximately follows an F distribution with k numerator degrees of freedom and $(s-k+1)$ denominator degrees of freedom.

Output Data Sets

You can use the Output Delivery System to create a SAS data set from any piece of PROC SURVEYFREQ output. For more information, see the section "ODS Table Names" on page 8045 and Example 97.3.

PROC SURVEYFREQ also provides an output data set that stores the replicate weights for BRR or jackknife variance estimation and an output data set that stores the jackknife coefficients for jackknife variance estimation.

Replicate Weight Output Data Set

If you specify the OUTWEIGHTS= *method-option* for VARMETHOD=BRR or VARMETHOD=JACKKNIFE, PROC SURVEYFREQ stores the replicate weights in an output data set. The OUTWEIGHTS= output data set contains all observations from the DATA= input data set that are valid (used in the analysis). A valid observation must have a positive value of the WEIGHT variable. A valid observations must also have nonmissing values of the STRATA and CLUSTER variables unless you specify the MISSING option in the PROC SURVEYFREQ statement. For information about valid observations, see the section "Data Summary Table" on page 8038.

The OUTWEIGHTS= data set contains the following variables:

- all variables in the DATA= input data set
- RepWt_1, RepWt_2, . . ., RepWt_n, which are the replicate weight variables, where n is the total number of replicates in the analysis

Each replicate weight variable contains the replicate weights for the corresponding replicate. Replicate weights equal zero for those observations not included in the replicate.

After the procedure creates and stores replicate weights for a particular input data set and survey design, you can use them again in subsequent analyses, either in PROC SURVEYFREQ or in another survey procedure. You use a REPWEIGHTS statement to provide replicate weights to the procedure.

Jackknife Coefficient Output Data Set

If you specify the OUTJKCOEFS= *method-option* for VARMETHOD=JACKKNIFE, PROC SURVEYFREQ stores the jackknife coefficients in an output data set. The OUTJKCOEFS= output data set contains one observation for each replicate. The OUTJKCOEFS= data set contains the following variables:

- Replicate, which is the replicate number for the jackknife coefficient
- JKCoefficient, which is the jackknife coefficient
- DonorStratum, which is the stratum of the PSU that was deleted to construct the replicate, if you specify a STRATA statement

After the procedure creates jackknife coefficients for a particular input data set and survey design, you can use the OUTJKCOEFS= *method-option* to store these coefficients and then use them again in subsequent analyses, either in PROC SURVEYFREQ or in another survey procedure. You use the JKCOEFS= option in the REPWEIGHTS statement to provide jackknife coefficients for the procedure.

Displayed Output

Data Summary Table

The “Data Summary” table provides information about the input data set and the sample design. PROC SURVEYFREQ displays this table unless you specify the NOSUMMARY option in the PROC SURVEYFREQ statement.

The “Data Summary” table displays the total number of valid observations. To be considered *valid*, an observation must have a nonmissing, positive sampling weight value if you specify a WEIGHT statement. If you do not specify the MISSING option, a valid observation must also have nonmissing values for all STRATA and CLUSTER variables. The number of valid observations can differ from the number of nonmissing observations for an individual table request, which the procedure displays in the frequency or crosstabulation tables. For more information, see the section “Missing Values” on page 8001.

PROC SURVEYFREQ displays the following information in the “Data Summary” table:

- Number of Strata, if you specify a STRATA statement
- Number of Clusters, if you specify a CLUSTER statement
- Number of Observations, which is the total number of valid observations
- Sum of Weights, which is the sum over all valid observations, if you specify a WEIGHT or REPWEIGHTS statement

Stratum Information Table

If you specify the LIST option in the STRATA statement, PROC SURVEYFREQ displays a “Stratum Information” table. This table provides the following information for each stratum:

- Stratum Index, which is a sequential stratum identification number
- STRATA variables, which list the levels of STRATA variables for the stratum
- Number of Observations, which is the number of valid observations in the stratum
- Population Total for the stratum, if you specify the TOTAL= option
- Sampling Rate for the stratum, if you specify the TOTAL= or RATE= option. If you specify the TOTAL= option, the sampling rate is based on the number of valid observations in the stratum.
- Number of Clusters, which is the number of clusters in the stratum, if you specify a CLUSTER statement

Variance Estimation Table

If you specify the VARMETHOD=BRR, VARMETHOD=JACKKNIFE, or NOMCAR option in the PROC SURVEYFREQ statement, the procedure displays a “Variance Estimation” table. If you do not specify any of these options, the procedure creates a “Variance Estimation” table but does not display it. You can store this nondisplayed table in an output data set by using the Output Delivery System (ODS). For more information, see the section “ODS Table Names” on page 8045.

The “Variance Estimation” table provides the following information:

- Method, which is the variance estimation method—Taylor Series, Balanced Repeated Replication, or Jackknife
- Replicate Weights input data set name, if you use a REPWEIGHTS statement to provide replicate weights
- Number of Replicates, if you specify VARMETHOD=BRR or VARMETHOD=JACKKNIFE
- Hadamard Data Set name, if you specify the HADAMARD= *method-option* for VARMETHOD=BRR
- Fay Coefficient, if you specify the FAY *method-option* for VARMETHOD=BRR
- Missing Levels Included (MISSING), if you specify the MISSING option
- Missing Levels Included (NOMCAR), if you specify the NOMCAR option

Hadamard Matrix

If you specify the PRINTH *method-option* for VARMETHOD=BRR, PROC SURVEYFREQ displays the Hadamard matrix that it uses to construct replicates for BRR variance estimation. If you provide a Hadamard matrix by specifying the HADAMARD= *method-option* for VARMETHOD=BRR but the procedure does not use the entire matrix, the procedure displays only the rows and columns that are actually used to construct replicates.

One-Way Frequency Tables

PROC SURVEYFREQ displays one-way frequency tables for all one-way table requests in the TABLES statements, unless you specify the NOPRINT option in the TABLES statement. A one-way table shows the sample frequency distribution of a single variable, and provides estimates for its population distribution in terms of totals and proportions.

If you request a one-way table without specifying options, PROC SURVEYFREQ displays the following information for each level of the variable:

- Frequency count, which is the number of sample observations in the level
- Weighted Frequency, which estimates the population total for the level
- Standard Deviation of Weighted Frequency
- Percent, which estimates the population proportion for the level
- Standard Error of Percent

The one-way table displays weighted frequencies if your analysis includes a WEIGHT or REPWEIGHTS statement, or if you specify the WTFREQ option in the TABLES statement.

The one-way table also displays the Frequency Missing, which is the number of observations with missing values.

You can suppress the frequency counts by specifying the NOFREQ option in the TABLES statement. Also, the NOWT option suppresses the weighted frequencies and their standard deviations. The NOPERCENT option suppresses the percentages and their standard errors. The NOSTD option suppresses the standard errors of the percentages and the standard deviations of the weighted frequencies. The NOTOTAL option suppresses the total row of the one-way table.

PROC SURVEYFREQ optionally displays the following information in a one-way table:

- Variance of Weighted Frequency, if you specify the VARWT option
- Confidence Limits for Weighted Frequency, if you specify the CLWT option
- Coefficient of Variation for Weighted Frequency, if you specify the CVWT option
- Test Percent, if you specify the TESTP= option
- Variance of Percent, if you specify the VAR option
- Confidence Limits for Percent, if you specify the CL option
- Coefficient of Variation for Percent, if you specify the CV option
- Design Effect for Percent, if you specify the DEFF option

Crosstabulation Tables

PROC SURVEYFREQ displays all table requests in the TABLES statements, unless you specify the NOPRINT option in the TABLES statement. For two-way to multiway crosstabulation tables, the values of the last variable in the table request form the table columns. The values of the next-to-last variable form the rows. Each level (or combination of levels) of the other variables forms one layer. PROC SURVEYFREQ produces a separate two-way crosstabulation table for each layer of a multiway table.

For each layer, the crosstabulation table displays the row and column variable names and values (levels). Each two-way table lists levels of the column variable within each level of the row variable.

By default, the procedure displays all levels of the column variable within each level of the row variables, including any column variable levels with zero frequency for that row. For multiway tables, the procedure displays all levels of the row variable for each layer of the table by default, including any row levels with zero frequency for that layer. You can suppress the display of zero frequency levels by specifying the NOSPARSE option.

If you request a crosstabulation table without specifying options, the table displays the following information for each combination of variable levels (table cell):

- Frequency, which is the number of sample observations in the table cell
- Weighted Frequency, which estimates the population total for the table cell
- Standard Deviation of Weighted Frequency
- Percent, which estimates the population proportion for the table cell
- Standard Error of Percent

The two-way table displays weighted frequencies if your analysis includes a WEIGHT or REPWEIGHTS statement, or if you specify the WTFREQ option in the TABLES statement.

The two-way table also displays the Frequency Missing, which is the number of observations with missing values.

You can suppress the frequency counts by specifying the NOFREQ option in the TABLES statement. Also, the NOWT option suppresses the weighted frequencies and their standard deviations. The NOPERCENT option suppresses all percentages and their standard errors. The NOCELLPERCENT option suppresses overall cell percentages and their standard errors, but displays any other percentages (and standard errors) that you request, such as row or column percentages. The NOSTD option suppresses the standard errors of the percentages and the standard deviations of the weighted frequencies. The NOTOTAL option suppresses the row totals and column totals, as well as the overall total.

PROC SURVEYFREQ optionally displays the following information in a two-way table:

- Expected Weighted Frequency, if you specify the EXPECTED option
- Deviation from Expected Weighted Frequency, if you specify the DEVIATION option

- Pearson Residual, if you specify the PEARSONRES option
- Cell Chi-Square, if you specify the CELLCHI2 option
- Variance of Weighted Frequency, if you specify the VARWT option
- Confidence Limits for Weighted Frequency, if you specify the CLWT option
- Coefficient of Variation for Weighted Frequency, if you specify the CVWT option
- Variance of Percent, if you specify the VAR option
- Confidence Limits for Percent, if you specify the CL option
- Coefficient of Variation for Percent, if you specify the CV option
- Design Effect for Percent, if you specify the DEFF option
- Row Percent, which estimates the population proportion of the row total, if you specify the ROW option
- Standard Error of Row Percent, if you specify the ROW option
- Variance of Row Percent, if you specify the VAR option and the ROW option
- Confidence Limits for Row Percent, if you specify the CL option and the ROW option
- Coefficient of Variation for Row Percent, if you specify the CV option and the ROW option
- Design Effect for Row Percent, if you specify the ROW(DEFF) option
- Column Percent, which estimates the population proportion of the column total, if you specify the COLUMN option
- Standard Error of Column Percent, if you specify the COLUMN option
- Variance of Column Percent, if you specify the VAR option and the COLUMN option
- Confidence Limits for Column Percent, if you specify the CL option and the COLUMN option
- Coefficient of Variation for Column Percent, if you specify the CV option and the COLUMN option
- Design Effects for Column Percent, if you specify the COLUMN(DEFF) option

Covariance Matrices of Estimates

If you specify the COV option, PROC SURVEYFREQ displays the covariance matrix of the cell total frequency estimates. If you specify the COVP option, PROC SURVEYFREQ displays the covariance matrix of the proportion estimates.

Statistical Tests

If you specify the CHISQ option for the Rao-Scott chi-square test or the LRCHISQ option for the Rao-Scott likelihood ratio chi-square test, PROC SURVEYFREQ displays the following information:

- Pearson Chi-Square, if you specify the CHISQ option
- Likelihood Ratio Chi-Square, if you specify the LRCHISQ option
- Design Correction
- Rao-Scott Chi-Square, by default or if you specify the FIRSTORDER option
- First-Order Chi-Square, if you specify the SECONDORDER option
- Second-Order Chi-Square, if you specify the SECONDORDER option
- DF, which is the degrees of freedom for the chi-square test
- Pr > ChiSq, which is the *p*-value for the chi-square test
- F Value
- Num DF, which is the numerator degrees of freedom for F
- Den DF, which is the denominator degrees of freedom for F
- Pr > F, which is the *p*-value for the *F* test

If you specify the WCHISQ option for the Wald chi-square test or the WLLCHISQ option for the Wald log-linear chi-square test, PROC SURVEYFREQ displays the following information:

- Wald Chi-Square, if you specify the WCHISQ option
- Wald Log-Linear Chi-Square, if you specify the WLLCHISQ option
- F Value
- Num DF, which is the numerator degrees of freedom for F
- Den DF, which is the denominator degrees of freedom for F
- Pr > F, which is the *p*-value for the *F* test
- Adjusted F Value, for tables larger than 2×2
- Num DF, which is the numerator degrees of freedom for Adjusted F
- Den DF, which is the denominator degrees of freedom for Adjusted F
- Pr > Adj F, which is the *p*-value for the Adjusted *F* test

Risks and Risk Difference

If you specify the RISK option in the TABLES statement for a 2 × 2 table, PROC SURVEYFREQ displays "Column 1 Risk Estimates" and "Column 2 Risk Estimates" tables. You can display only column 1 or column 2 risks by specifying the RISK1 or RISK2 option, respectively.

The "Risk Estimates" table displays the following information for Row 1, Row 2, Total, and Difference:

- Row, which identifies the risk as Row 1, Row 2, Total, or Difference
- Risk estimate
- Standard Error
- Confidence Limits

In the "Column 1 Risk Estimates" table, the row 1 risk is the column 1 percentage of row 1. The row 2 risk is the column 1 percentage of row 2, and the total risk is the column 1 percentage of the entire table. The risk difference is the row 1 risk minus the row 2 risk. In the "Column 2 Risk Estimates" table, these computations are based on column 2.

Odds Ratio and Relative Risks

If you specify the OR option in the TABLES statement for a 2 × 2 table, PROC SURVEYFREQ displays the "Odds Ratio" table. This table includes the following information:

- Statistic, which identifies the statistic as the Odds Ratio, the Column 1 Relative Risk, or the Column 2 Relative Risk
- Estimate
- Confidence Limits

Kappa Statistics

If you specify the AGREE, KAPPA, or WTKAPPA option in the TABLES statement for a square table, PROC SURVEYFREQ displays the "Kappa Statistics" table. This table includes the following information:

- Statistic, which identifies the statistic as the Simple Kappa Coefficient or the Weighted Kappa Coefficient
- Estimate
- Standard Error
- Confidence Limits

Kappa Weights

If you specify the AGREE(PRINTKWTS) or WTKAPPA(PRINTKWTS) option for a square table whose dimension is greater than 2, PROC SURVEYFREQ displays the “Kappa Weights” table. This table provides the matrix of kappa agreement weights that the procedure uses to compute the weighted kappa coefficient. The matrix contains an agreement weight for each pair of column variable levels.

ODS Table Names

PROC SURVEYFREQ assigns a name to each table that it creates. You can use these names to refer to tables when you use the Output Delivery System (ODS) to select tables and create output data sets. For more information about ODS, see Chapter 20, “Using the Output Delivery System.” See Example 97.3 for examples of storing PROC SURVEYFREQ tables as output data sets.

Table 97.9 lists the ODS table names together with their descriptions and the options required to produce the tables.

Table 97.9 ODS Tables Produced by PROC SURVEYFREQ

ODS Table Name	Description	Statement	Option
ChiSq	Chi-square test	TABLES	CHISQ
ChiSq1	Modified chi-square test	TABLES	CHISQ(MODIFIED)
Cov	Covariances of frequencies	TABLES	COV
CovP	Covariances of proportions	TABLES	COVP
CrossTabs	Crosstabulation table	TABLES	n-way table request, $n > 1$
HadamardMatrix	Hadamard matrix	PROC	VARMETHOD=BRR(PRINTH)
Kappa	Kappa coefficients	TABLES	AGREE, KAPPA, or WTKAPPA ($r \times r$ table)
KappaWeights	Kappa agreement weights	TABLES	WTKAPPA(PRINTKWTS) ($r \times r$ table, $r > 2$)
LRChiSq	Likelihood ratio test	TABLES	LRCHISQ
LRChiSq1	Modified likelihood ratio test	TABLES	LRCHISQ(MODIFIED)
OddsRatio	Odds ratio and relative risks	TABLES	OR (2×2 table)
OneWay	One-way frequency table	PROC or TABLES	No TABLES statement One-way table request
Risk1	Column 1 risk estimates	TABLES	RISK or RISK1 (2×2 table)
Risk2	Column 2 risk estimates	TABLES	RISK or RISK2 (2×2 table)
StrataInfo	Stratum information	STRATA	LIST
Summary	Data summary	PROC	Default
TableSummary	Table summary (not displayed)	TABLES	Default
VarianceEstimation	Variance estimation	PROC	VARMETHOD=BRR, VARMETHOD=JACKKNIFE, or NOMCAR
WChiSq	Wald chi-square test	TABLES	WCHISQ (two-way table)
WLLChiSq	Wald log-linear chi-square test	TABLES	WLLCHISQ (two-way table)

ODS Graphics

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 21, "Statistical Graphics Using ODS."

Before you create graphs, ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 606 in Chapter 21, "Statistical Graphics Using ODS."

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section "A Primer on ODS Statistical Graphics" on page 605 in Chapter 21, "Statistical Graphics Using ODS."

When ODS Graphics is enabled, you can request specific plots by specifying the PLOTS= option in the TABLES statement. To produce a weighted frequency plot or mosaic plot, you must specify the WTFREQPLOT or MOSAICPLOT *plot-request* in the PLOTS= option, or you must specify the PLOTS=ALL option. By default, PROC SURVEYFREQ produces all other plots that are associated with the analyses that you request in the TABLES statement. You can suppress default plots and request specific plots by using the PLOTS(ONLY)= option. For more information, see the description of the PLOTS= option.

PROC SURVEYFREQ assigns a name to each graph that it creates by using ODS Graphics. You can use these names to refer to the graphs. Table 97.10 lists the names of the graphs that PROC SURVEYFREQ generates together with their descriptions, their PLOTS= options (*plot-requests*), and the TABLES statement options that are required to produce the graphs.

Table 97.10 ODS Graphs Produced by PROC SURVEYFREQ

ODS Graph Name	Description	PLOTS= Option	TABLES Statement Option
KappaPlot	Kappa plot	KAPPAPLOT	AGREE or KAPPA ($h \times r \times r$ table)
MosaicPlot	Mosaic plot	MOSAICPLOT	Two-way or multiway table request
ORPlot	Odds ratio plot	ODDSRATIOPLOT	OR ($h \times 2 \times 2$ table)
RelRiskPlot	Relative risk plot	RELRISKPLOT	OR ($h \times 2 \times 2$ table)
RiskDiffPlot	Risk difference plot	RISKDIFFPLOT	RISK ($h \times 2 \times 2$ table)
WtFreqPlot	Weighted frequency plot	WTFREQPLOT	Any table request
WtKappaPlot	Weighted kappa plot	WTKAPPAPLOT	AGREE or WTKAPPA ($h \times r \times r$ table, $r > 2$)

Examples: SURVEYFREQ Procedure

Example 97.1: Two-Way Tables

This example uses the SIS_Survey data set from the section "Getting Started: SURVEYFREQ Procedure" on page 7958. The data set contains results from a customer satisfaction survey for a student information system (SIS).

The following PROC SURVEYFREQ statements request a two-way table for Department by Response and customize the crosstabulation table display:

```
title 'Student Information System Survey';
proc surveyfreq data=SIS_Survey;
   tables  Department * Response / cv deff nowt nostd nototal;
   strata  State NewUser / list;
   cluster School;
   weight  SamplingWeight;
run;
```

The TABLES statement requests a two-way table of Department by Response. The CV option requests coefficients of variation for the percentage estimates. The DEFF option requests design effects for the percentage estimates. The NOWT option suppresses display of the weighted frequencies, and the NOSTD option suppresses display of standard errors for the estimates. The NOTOTAL option suppresses the row totals, column totals, and overall totals.

The STRATA, CLUSTER, and WEIGHT statements provide sample design information for the procedure, so that the analysis is done according to the sample design used for the survey. The STRATA statement names the variables State and NewUser, which identify the first-stage strata. The LIST option in the STRATA statement requests a "Stratum Information" table. The CLUSTER statement names the variable School, which identifies the clusters (primary sampling units). The WEIGHT statement names the sampling weight variable.

Output 97.1.1 displays the "Data Summary" and "Stratum Information" tables produced by PROC SURVEYFREQ. The "Stratum Information" table lists the six strata in the survey and shows the number of observations and the number of clusters (schools) in each stratum.

Output 97.1.1 Data Summary and Stratum Information

Student Information System Survey

The SURVEYFREQ Procedure

Data Summary	
Number of Strata	6
Number of Clusters	370
Number of Observations	1850
Sum of Weights	38899.6482

Stratum Information

Stratum Index	State	NewUser	Number of Obs	Number of Clusters
1	GA	Renewal Customer	315	63
2	GA	New Customer	355	71
3	NC	Renewal Customer	280	56
4	NC	New Customer	420	84
5	SC	Renewal Customer	210	42
6	SC	New Customer	270	54

Output 97.1.2 displays the two-way table of Department by Response. According to the TABLES statement options that are specified, this two-way table includes coefficients of variation and design effects for the percentage estimates, and it does not show the weighted frequencies or the standard errors of the estimates. It also does not show the row, column, and overall totals.

Output 97.1.2 Two-Way Table of Department by Response

Table of Department by Response

Department	Response	Frequency	Percent	CV for Percent	Design Effect
Faculty	**Very Unsatisfied**	209	13.4987	0.0865	2.1586
	Unsatisfied	203	13.0710	0.0868	2.0962
	Neutral	346	22.4127	0.0629	2.1157
	Satisfied	254	16.2006	0.0806	2.3232
	Very Satisfied	98	6.2467	0.1362	2.2842
Admin/Guidance	**Very Unsatisfied**	95	3.6690	0.1277	1.1477
	Unsatisfied	123	4.6854	0.1060	1.0211
	Neutral	235	9.1838	0.0700	0.9166
	Satisfied	201	7.7305	0.0756	0.8848
	Very Satisfied	86	3.3016	0.1252	0.9892

The following PROC SURVEYFREQ statements request a two-way table of Department by Response that includes row percentages, and also a Wald chi-square test of association between the two table variables:

```
title 'Student Information System Survey';
proc surveyfreq data=SIS_Survey nosummary;
   tables  Department * Response / row nowt wchisq;
   strata  State NewUser;
   cluster School;
   weight  SamplingWeight;
run;
```

Output 97.1.3 displays the two-way table. The row percentages show the distribution of Response for Department = ‘Faculty’ and for Department = ‘Admin/Guidance’. This is equivalent to a domain (subpopulation) analysis of Response, where the domains are Department = ‘Faculty’ and Department = ‘Admin/Guidance’.

Output 97.1.4 displays the Wald chi-square test of association between Department and Response. The Wald chi-square is 11.44, and the corresponding adjusted F value is 2.84 with a p-value of 0.0243. This indicates a significant association between department (faculty or admin/guidance) and satisfaction with the student information system.

Output 97.1.3 Table of Department by Response with Row Percentages

Student Information System Survey

The SURVEYFREQ Procedure

Table of Department by Response						
Department	**Response**	**Frequency**	**Percent**	**Std Err of Percent**	**Row Percent**	**Std Err of Row Percent**
Faculty	**Very Unsatisfied**	209	13.4987	1.1675	18.8979	1.6326
	Unsatisfied	203	13.0710	1.1350	18.2992	1.5897
	Neutral	346	22.4127	1.4106	31.3773	1.9705
	Satisfied	254	16.2006	1.3061	22.6805	1.8287
	Very Satisfied	98	6.2467	0.8506	8.7452	1.1918
	Total	1110	71.4297	0.1468	100.000	
Admin/Guidance	**Very Unsatisfied**	95	3.6690	0.4684	12.8419	1.6374
	Unsatisfied	123	4.6854	0.4966	16.3995	1.7446
	Neutral	235	9.1838	0.6430	32.1447	2.2300
	Satisfied	201	7.7305	0.5842	27.0579	2.0406
	Very Satisfied	86	3.3016	0.4133	11.5560	1.4466
	Total	740	28.5703	0.1468	100.000	
Total	**Very Unsatisfied**	304	17.1676	1.2872		
	Unsatisfied	326	17.7564	1.2712		
	Neutral	581	31.5965	1.5795		
	Satisfied	455	23.9311	1.4761		
	Very Satisfied	184	9.5483	0.9523		
	Total	1850	100.000			

Output 97.1.4 Wald Chi-Square Test

Wald Chi-Square Test	
Chi-Square	11.4454
F Value	2.8613
Num DF	4
Den DF	364
Pr > F	0.0234
Adj F Value	2.8378
Num DF	4
Den DF	361
Pr > Adj F	0.0243
Sample Size = 1850	

Example 97.2: Multiway Tables (Domain Analysis)

Continuing to use the SIS_Survey data set from the section "Getting Started: SURVEYFREQ Procedure" on page 7958, this example shows how to produce multiway tables. The following PROC SURVEYFREQ statements request a table of Department by SchoolType by Response for the student information system survey:

```
title 'Student Information System Survey';
proc surveyfreq data=SIS_Survey;
   tables  Department * SchoolType * Response
           SchoolType * Response;
   strata  State NewUser;
   cluster School;
   weight  SamplingWeight;
run;
```

The TABLES statement requests a multiway table with SchoolType as the row variable, Response as the column variable, and Department as the layer variable. This request produces a separate two-way table of SchoolType by Response for each level of the variable Department. The TABLES statement also requests a two-way table of SchoolType by Response, which totals the multiway table over both levels of Department. As in the previous examples, the STRATA, CLUSTER, and WEIGHT statements provide sample design information, so that the analysis will be done according to the design used for this survey.

Output 97.2.1 displays the multiway table produced by PROC SURVEYFREQ, which includes a table of SchoolType by Response for Department = 'Faculty' and for Department = 'Admin/Guidance'. This is equivalent to a domain (subpopulation) analysis of SchoolType by Response, where the domains are Department = 'Faculty' and Department = 'Admin/Guidance'.

Output 97.2.1 Multiway Table of Department by SchoolType by Response

Student Information System Survey

The SURVEYFREQ Procedure

Table of SchoolType by Response
Controlling for Department=Faculty

SchoolType	Response	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
Middle School	Very Unsatisfied	74	1846	301.22637	6.6443	1.0838
	Unsatisfied	78	1929	283.11476	6.9428	1.0201
	Neutral	130	3289	407.80855	11.8369	1.4652
	Satisfied	113	2795	368.85087	10.0597	1.3288
	Very Satisfied	55	1378	261.63311	4.9578	0.9411
	Total	450	11237	714.97120	40.4415	2.5713
High School	Very Unsatisfied	135	3405	389.42313	12.2536	1.3987
	Unsatisfied	125	3155	384.56734	11.3563	1.3809
	Neutral	216	5429	489.37826	19.5404	1.7564
	Satisfied	141	3507	417.54773	12.6208	1.5040
	Very Satisfied	43	1052	221.59367	3.7874	0.7984
	Total	660	16549	719.61536	59.5585	2.5713
Total	Very Unsatisfied	209	5251	454.82598	18.8979	1.6326
	Unsatisfied	203	5085	442.39032	18.2992	1.5897
	Neutral	346	8718	550.81735	31.3773	1.9705
	Satisfied	254	6302	507.01711	22.6805	1.8287
	Very Satisfied	98	2430	330.97602	8.7452	1.1918
	Total	1110	27786	119.25529	100.000	

Table of SchoolType by Response
Controlling for Department=Admin/Guidance

SchoolType	Response	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
Middle School	Very Unsatisfied	42	649.43427	133.06194	5.8435	1.1947
	Unsatisfied	31	460.35557	100.80158	4.1422	0.9076
	Neutral	104	1568	186.99946	14.1042	1.6804
	Satisfied	84	1269	165.71127	11.4142	1.4896
	Very Satisfied	39	574.93878	110.37243	5.1732	0.9942
	Total	300	4521	287.86832	40.6774	2.5801
High School	Very Unsatisfied	53	777.77725	136.41869	6.9983	1.2285
	Unsatisfied	92	1362	175.40862	12.2573	1.5806
	Neutral	131	2005	212.34804	18.0404	1.8990
	Satisfied	117	1739	190.07798	15.6437	1.7118
	Very Satisfied	47	709.37033	126.54394	6.3828	1.1371
	Total	440	6593	288.92483	59.3226	2.5801
Total	Very Unsatisfied	95	1427	182.28132	12.8419	1.6374
	Unsatisfied	123	1823	193.43045	16.3995	1.7446
	Neutral	235	3572	250.22739	32.1447	2.2300
	Satisfied	201	3007	226.82311	27.0579	2.0406
	Very Satisfied	86	1284	160.83434	11.5560	1.4466
	Total	740	11114	60.78850	100.000	

Example 97.3: Output Data Sets

PROC SURVEYFREQ uses the Output Delivery System (ODS) to create output data sets. This is a departure from older SAS procedures that provide OUTPUT statements for similar functionality. By using ODS, you can create a SAS data set from any piece of PROC SURVEYFREQ output. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

When selecting tables for ODS output data sets, you refer to tables by their ODS table names. Each table created by PROC SURVEYFREQ is assigned a name. See the section "ODS Table Names" on page 8045 for a list of the table names provided by PROC SURVEYFREQ.

To save the one-way table of Response from Figure 97.3 in an output data set, use an ODS OUTPUT statement as follows:

```
proc surveyfreq data=SIS_Survey;
   tables  Response / cl nowt;
   ods output OneWay=ResponseTable;
   strata  State NewUser;
   cluster School;
   weight  SamplingWeight;
run;
```

Output 97.3.1 displays the output data set ResponseTable, which contains the one-way table of Response. This data set has six observations, and each of these observations corresponds to a row of the one-way table. The first five observations correspond to the five levels of Response, as they are ordered in the one-way table display, and the last observation corresponds to the overall total, which is the last row of the one-way table. The data set ResponseTable includes a variable corresponding to each column of the one-way table. For example, the variable Percent contains the percentage estimates, and the variables LowerCL and UpperCL contain the lower and upper confidence limits for the percentage estimates.

Output 97.3.1 ResponseTable Output Data Set

Obs	Table	Response	Frequency	Percent	StdErr	LowerCL	UpperCL
1	Table Response	Very Unsatisfied	304	17.1676	1.2872	14.6364	19.6989
2	Table Response	Unsatisfied	326	17.7564	1.2712	15.2566	20.2562
3	Table Response	Neutral	581	31.5965	1.5795	28.4904	34.7026
4	Table Response	Satisfied	455	23.9311	1.4761	21.0285	26.8338
5	Table Response	Very Satisfied	184	9.5483	0.9523	7.6756	11.4210
6	Table Response	.	1850	100.000	_	_	_

PROC SURVEYFREQ also creates a table summary that is not displayed. Some of the information in this table is similar to that contained in the "Data Summary" table, but the table summary describes the data that are used to analyze the specified table, while the data summary describes the entire input data set. Due to missing values, for example, the number of observations (or strata or clusters) used to analyze a particular table can differ from the number of observations (or strata or clusters) reported for the input data set in the "Data Summary" table. For more information, see the section "Missing Values" on page 8001. If you request confidence limits, the "Table Summary" table also contains the degrees of freedom and the t-value used to compute the confidence limits.

The following statements store the nondisplayed “Table Summary” table in the output data set ResponseSummary:

```
proc surveyfreq data=SIS_Survey;
   tables  Response / cl nowt;
   ods output TableSummary=ResponseSummary;
   strata  State NewUser;
   cluster School;
   weight  SamplingWeight;
run;
```

Output 97.3.2 displays the output data set ResponseSummary.

Output 97.3.2 ResponseSummary Output Data Set

Obs	Table	Number of Observations	Number of Strata	Number of Clusters	Degrees of Freedom	t Percentile
1	Table Response	1850	6	370	364	1.966503

References

Agresti, A. (2007), *An Introduction to Categorical Data Analysis*, 2nd Edition, New York: John Wiley & Sons.

Agresti, A. (2013), *Categorical Data Analysis*, 3rd Edition, Hoboken, NJ: John Wiley & Sons.

Agresti, A. and Coull, B. A. (1998), “Approximate Is Better Than ‘Exact’ for Interval Estimation of Binomial Proportions,” *American Statistician*, 52, 119–126.

Bedrick, E. J. (1983), “Adjusted Chi-Squared Tests for Cross-Classified Tables of Survey Data,” *Biometrika*, 70, 591–596.

Brick, J. M. and Kalton, G. (1996), “Handling Missing Data in Survey Research,” *Statistical Methods in Medical Research*, 5, 215–238.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001), “Interval Estimation for a Binomial Proportion,” *Statistical Science*, 16, 101–133.

Cicchetti, D. V. and Allison, T. (1971), “A New Procedure for Assessing Reliability of Scoring EEG Sleep Recordings,” *American Journal of EEG Technology*, 11, 101–109.

Clopper, C. J. and Pearson, E. S. (1934), “The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial,” *Biometrika*, 26, 404–413.

Cochran, W. G. (1977), *Sampling Techniques*, 3rd Edition, New York: John Wiley & Sons.

Cohen, J. (1960), “A Coefficient of Agreement for Nominal Scales,” *Educational and Psychological Measurement*, 20, 37–46.

Curtin, L. R., Kruszon-Moran, D., Carroll, M., and Li, X. (2006), "Estimation and Analytic Issues for Rare Events in NHANES," in *Proceedings of the Survey Research Methods Section*, 2893–2903, Alexandria, VA: American Statistical Association.

Dippo, C. S., Fay, R. E., and Morganstein, D. H. (1984), "Computing Variances from Complex Samples with Replicate Weights," in *Proceedings of the Survey Research Methods Section*, 489–494, Alexandria, VA: American Statistical Association.

Fay, R. E. (1989), "Theory and Application of Replicate Weighting for Variance Calculations," in *Proceedings of the Survey Research Methods Section*, 212–217, Alexandria, VA: American Statistical Association.

Felligi, I. P. (1980), "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples," *Journal of the American Statistical Association*, 75, 261–268.

Fienberg, S. E. (1980), *The Analysis of Cross-Classified Categorical Data*, 2nd Edition, Cambridge, MA: MIT Press.

Fleiss, J. L. (1981), *Statistical Methods for Rates and Proportions*, 2nd Edition, New York: John Wiley & Sons.

Fleiss, J. L. and Cohen, J. (1973), "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability," *Educational and Psychological Measurement*, 33, 613–619.

Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969), "Large-Sample Standard Errors of Kappa and Weighted Kappa," *Psychological Bulletin*, 72, 323–327.

Fleiss, J. L., Levin, B., and Paik, M. C. (2003), *Statistical Methods for Rates and Proportions*, 3rd Edition, Hoboken, NJ: John Wiley & Sons.

Friendly, M. (2000), *Visualizing Categorical Data*, Cary, NC: SAS Institute Inc.

Fuller, W. A. (1975), "Regression Analysis for Sample Survey," *Sankhyā, Series C*, 37, 117–132.

Fuller, W. A., Kennedy, W. J., Schnell, D., Sullivan, G., and Park, H. J. (1989), *PC CARP*, Ames: Iowa State University Statistical Laboratory.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953), *Sample Survey Methods and Theory*, volume 1 and 2, New York: John Wiley & Sons.

Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. (1980), *SUPER CARP*, Ames: Iowa State University Statistical Laboratory.

Judkins, D. R. (1990), "Fay's Method for Variance Estimation," *Journal of Official Statistics*, 6, 223–239.

Kalton, G. (1983), *Introduction to Survey Sampling*, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-035, Beverly Hills, CA: Sage Publications.

Kalton, G. and Kasprzyk, D. (1986), "The Treatment of Missing Survey Data," *Survey Methodology*, 12, 1–16.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons.

Koch, G. G., Freeman, D. H., and Freeman, J. L. (1975), "Strategies in the Multivariate Analysis of Data from Complex Surveys," *International Statistical Review*, 43, 59–78.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., and Lehnen, R. G. (1977), "A General Methodology for the Analysis of Experiments with Repeated Measurement of Categorical Data," *Biometrics*, 33, 133–158.

Korn, E. L. and Graubard, B. I. (1990), "Simultaneous Testing with Complex Survey Data: Use of Bonferroni *t*-Statistics," *American Statistician*, 44, 270–276.

Korn, E. L. and Graubard, B. I. (1998), "Confidence Intervals for Proportions with Small Expected Number of Positive Counts Estimated from Survey Data," *Survey Methodology*, 24, 193–201.

Korn, E. L. and Graubard, B. I. (1999), *Analysis of Health Surveys*, New York: John Wiley & Sons.

Lee, E. S., Forthofer, R. N., and Lorimor, R. J. (1989), *Analyzing Complex Survey Data*, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-071, Beverly Hills, CA: Sage Publications.

Leemis, L. M. and Trivedi, K. S. (1996), "A Comparison of Approximate Interval Estimators for the Bernoulli Parameter," *American Statistician*, 50, 63–68.

Levy, P. S. and Lemeshow, S. (1999), *Sampling of Populations: Methods and Applications*, 3rd Edition, New York: John Wiley & Sons.

Lohr, S. L. (2010), *Sampling: Design and Analysis*, 2nd Edition, Boston: Brooks/Cole.

Nathan, G. (1975), "Tests for Independence in Contingency Tables from Stratified Samples," *Sankhyā, Series C*, 37, 77–87.

Newcombe, R. G. (1998), "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods," *Statistics in Medicine*, 17, 857–872.

Rao, J. N. K. and Scott, A. J. (1979), "Chi-Squared Tests for Analysis of Categorical Data from Complex Surveys," in *Proceedings of the Survey Research Methods Section*, 58–66, Washington, DC: American Statistical Association.

Rao, J. N. K. and Scott, A. J. (1981), "The Analysis of Categorical Data from Complex Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables," *Journal of the American Statistical Association*, 76, 221–230.

Rao, J. N. K. and Scott, A. J. (1984), "On Chi-Squared Tests for Multiway Contingency Tables with Cell Properties Estimated from Survey Data," *Annals of Statistics*, 12, 46–60.

Rao, J. N. K. and Scott, A. J. (1987), "On Simple Adjustments to Chi-Square Tests with Survey Data," *Annals of Statistics*, 15, 385–397.

Rao, J. N. K. and Shao, J. (1996), "On Balanced Half-Sample Variance Estimation in Stratified Random Sampling," *Journal of the American Statistical Association*, 91, 343–348.

Rao, J. N. K. and Shao, J. (1999), "Modified Balanced Repeated Replication for Complex Survey Data," *Biometrika*, 86, 403–415.

Rao, J. N. K. and Thomas, D. R. (1989), "Chi-Squared Tests for Contingency Tables," in C. J. Skinner, D. Holt, and T. M. F. Smith, eds., *Analysis of Complex Surveys*, 89–114, New York: John Wiley & Sons.

Särndal, C. E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, 2, 110–114.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2000), *Categorical Data Analysis Using the SAS System*, 2nd Edition, Cary, NC: SAS Institute Inc.

Sukasih, A. and Jang, D. (2005), "An Application of Confidence Interval Methods for Small Proportions in the Health Care Survey of DoD Beneficiaries," in *Proceedings of the Survey Research Methods Section*, 3608–3612, Alexandria, VA: American Statistical Association.

Thomas, D. R. and Rao, J. N. K. (1984), "A Monte Carlo Study of Exact Levels of Goodness-of-Fit Statistics under Cluster Sampling," in *Proceedings of the Survey Research Methods Section*, 207–211, Alexandria, VA: American Statistical Association.

Thomas, D. R. and Rao, J. N. K. (1985), "On the Power of Some Goodness-of-Fit Tests under Cluster Sampling," in *Proceedings of the Survey Research Methods Section*, 291–296, Alexandria, VA: American Statistical Association.

Thomas, D. R. and Rao, J. N. K. (1987), "Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics under Cluster Sampling," *Journal of the American Statistical Association*, 82, 630–636.

Thomas, D. R., Singh, A. C., and Roberts, G. R. (1996), "Tests of Independence on Two-Way Tables under Cluster Sampling: An Evaluation," *International Statistical Review*, 64, 295–311.

Wald, A. (1943), "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large," *Transactions of the American Mathematical Society*, 54, 426–482.

Wilson, E. B. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, 22, 209–212.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.

Woodruff, R. S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association*, 66, 411–414.

Subject Index

Syntax Index